University of London

Applied Optics Section
Department of Physics
Imperial College of Science and Technology

-

"The spatial coherence of thermal light sources"

A thesis submitted for the
degree of Doctor of Philosophy
by
John Denton Armitage, D.I.C.

February 1968

ABSTRACT

It is postulated that two points on a thermal light source, if they are sufficiently close to one another, may exhibit some degree of coherence between their radiations. The treatment of partial coherence due to Hopkins, in which a thermal source is considered to be perfectly incoherent, is extended to include the possibility of small regions of spatial coherence in the source. Without assuming any specific source emission mechanism other than that the coherence time of the radiation is much shorter than the integrating time of the detector used, the following rules, which are based on physical arguments rather than mathematical idealizations, are shown to apply:

i) Fourier spectral components of the same frequency coming from the same element of the source may be considered to be perfectly coherent.

ii) Fourier spectral components of the same frequency coming from different elements of the source may be considered to be partially coherent, the degree of partial coherence being determined by the properties of the source.

iii) Fourier spectral components of different frequencies may be considered to be completely incoherent.

It is shown that the assumption ordinarily made for a thermal source, that in rule ii) the different source elements are assumed to be perfectly incoherent, is justified for the usual situation in which they are spatially unresolved by an observing system.

Formulae are developed which describe the power at a point, and the coherence between points, resulting at an area illuminated by such a source. For a model source Gaussian in both radiance and micro-coherence, a computation shows that in order to obtain reasonable experimental measurements of source micro-coherence characteristics, the source must be extremely small.

TABLE OF CONTENTS

If anything physical comes out of mathematics,
it must have been put in in another form.

Bridgman (1927, p. 169)

INTRODUCTION

Until recently, it has generally been assumed that light disturbances (even if exactly the same frequency) emitted by two different atoms of a thermal light source, or that light disturbances of very slightly differing frequencies (even if emitted by the same atom), are completely incoherent - that is, they can not interfere to cancel or augment each other. This concept is most strictly stated by Dirac (1958, p. 9) as "Each photon then interferes only with itself. Interference between two different photons never occurs."

The development of classical coherence theory (Hopkins 1967; Born and Wolf 1959) has led to the concept of a degree of partial coherence between two points in a wave field, formulated in terms of the degree of correlation of the phases of the disturbances over a specified time. In the classical treatment, the partial degree of coherence is considered to arise from the two field points receiving coherent light from any given point in the light source, but with this light being completely incoherent with light received from all other points in the source. Thus the concept of the mutual incoherence of different atoms in the source has been retained, partial coherence only existing in a radiation field determined by a source but not at the source itself.

Both experiments demonstrating interference between light from two separate lasers (Mandel, 1963), and the existence of phase correlation over regions within the laser itself, suggest that the assumption of complete incoherence between different elements in a thermal light source may not be justified.[x] Further, the coherence between different atoms stimulated to emit by the standing wave field of the laser, and the small but non-zero coherent component of radiation generated by stimulated emission in a black-body cavity (Heavens 1964, p. 3), both suggest that for source atoms sufficiently close together, the radiation emitted spontaneously from one atom could stimulate nearby atoms to emit radiation coherent with the stimulating radiation. This could lead to very small regions of coherence in an ordinary light source, or "micro-coherence."

Although this is a difficult problem in radiation physics, and standard theory does not so far provide a prediction of the degree of micro-coherence to be expected, it is nevertheless worthwhile to examine the problem from

[x] It should be noted that Mandel (1964) points out that from the point of view of quantum theory, Dirac's statement (quoted earlier) should not be interpreted to mean that the light in two independent laser beams can not interfere, because in laser radiation the "average photon occupation number per unit cell of phase space is appreciably greater than one," so "the two beams cannot meaningfully be described as being incoherent or statistically independent," and a photon can be regarded as being partly in both beams and thus "interferes only with itself."

the viewpoint of instrumental optics to see whether a degree of micro-coherence can be defined phenomenologically and measured experimentally. For the design of such an experiment, formulae are required which provide a basis for measurement in terms of a suitably specified micro-coherence of the source and the characteristics of the apparatus used.

Hopkins' treatment of partial coherence (Hopkins 1967, p. 210) formulated in terms of monochromatic complex amplitudes associated with the radiation disturbances, while not directly applicable, suggests a useful extension to treat the problem of the micro-coherent source. Hopkins' treatment uses the assumed perfect incoherence of a thermal source in an argument showing that the light intensity produced at any point by an extended source may be found by summing the intensities produced at any point by pure monochromatic waves of different frequencies assumed to emerge independently from each element of the source, and that the total intensity is then found by integrating these monochromatic intensities over the appropriate spectral range. This principle has been used to develop a theory of partial coherence for points in the wave field produced by a source, wherein the coherence factor is shown to determine the visibility obtaining in any interference pattern.

Hopkins develops the following three rules for a classically incoherent source:

i) Light beams of the same frequency coming from the same element of the source are perfectly coherent.

ii) Light beams of the same frequency coming from different elements of the source may be considered to be completely incoherent.

iii) Light beams of different frequency are completely incoherent.

In deriving these rules, it has been assumed that light disturbances are emitted entirely independently from two different atoms, in accordance with the classical assumptions regarding thermal light sources. In the next section it will be shown that these rules are indeed correct even should the source be micro-coherent rather than completely incoherent, provided that the regions of micro-coherence are well below being resolvable by the apparatus viewing the source. The theory based on a completely incoherent source will be extended to include the general case of a micro-coherent source viewed without any limitations as to resolution, and it will be shown that rule ii) given above is only a special case of the general rule:

ii') Light beams of the same frequency coming from different elements of the source may be considered to be partially coherent, the degree of partial

coherence being determined by the properties
of the source.

It must be emphasized that both in Hopkins' derivation of the three rules given above, and in the present extension of the theory to include the micro-coherent source, it has been assumed that the integration time of the photodetecting process is much greater than the coherence time of the light in question. This condition is well satisfied for all but laser radiation, which is therefore excluded from the present treatment. For the situation wherein the detecting integration time is comparable to the radiation coherence time, the detailed statistical characteristics of the radiation must be known and very different methods must be used to treat the problem.

Included in the following treatment of source coherence will also be a careful consideration of the physical validity of the application of Fourier transform methods to this type of problem. It will be shown that straightforward application of Fourier methods is physically not justifiable, because the infinitely extended Fourier sin/cos components usually used lead to the following contradictions:

1) The postulated incoherence of infinitely extended pure monofrequency disturbances of identical frequency, which by their very definition must be perfectly coherent.

2) The existence of a Fourier component, and therefore the existence of its ability to create an effect on a detector, at a time for which the disturbance producing the Fourier component does not itself exist. This implies the response of a detector in advance of the arrival of the signal, and of its continued response after the ringing transients (during the detector integration time) has ceased.

These questions will be examined in detail, and it will be shown that both a modification of the usual Fourier transform and consideration of the general statistical nature of the emission process is required. However, no specific statistical properties will be assumed for the radiation other than that of a coherence time.

MICRO-COHERENCE

Specification of the Micro-Coherence Factor

Let the scalar $S(t,\underline{x})$ be the real temporal physical field disturbance at a point in the source having vector coordinate $\underline{x}$. (Fig. 1). The vector $\underline{x}$ may be either two or three-dimensional corresponding to either a surface or volume source. $S'(t,\underline{x}')$ will be used to denote the disturbance at time t at the point $\underline{x}'$ of a detector which receives the light. $S'(t,\underline{x}')$ may be written in terms of its Fourier spectrum, that is

$$S'(t,\underline{x}') = \int\limits_{\substack{-\infty \\ \nu}}^{\infty} s'(\nu,\underline{x}')\, e^{i2\pi\nu t}\, d\nu \ , \tag{1}$$

where

$$s'(\nu,\underline{x}') = \int\limits_{\substack{-\infty \\ t}}^{\infty} S'(t,\underline{x}')\, e^{-i2\pi\nu t}\, dt \ , \tag{2}$$

and similarly for the disturbance at the source. There are two problems with this standard mathematical representation. First, suppose the disturbance at the detector is examined at time $t = t'$. Using the above relations implies that, at a given moment $t = t'$, the Fourier spectrum can be determined by applying to $S'(t,\underline{x})$ an integral operator with limits $t = \pm\infty$. This is not

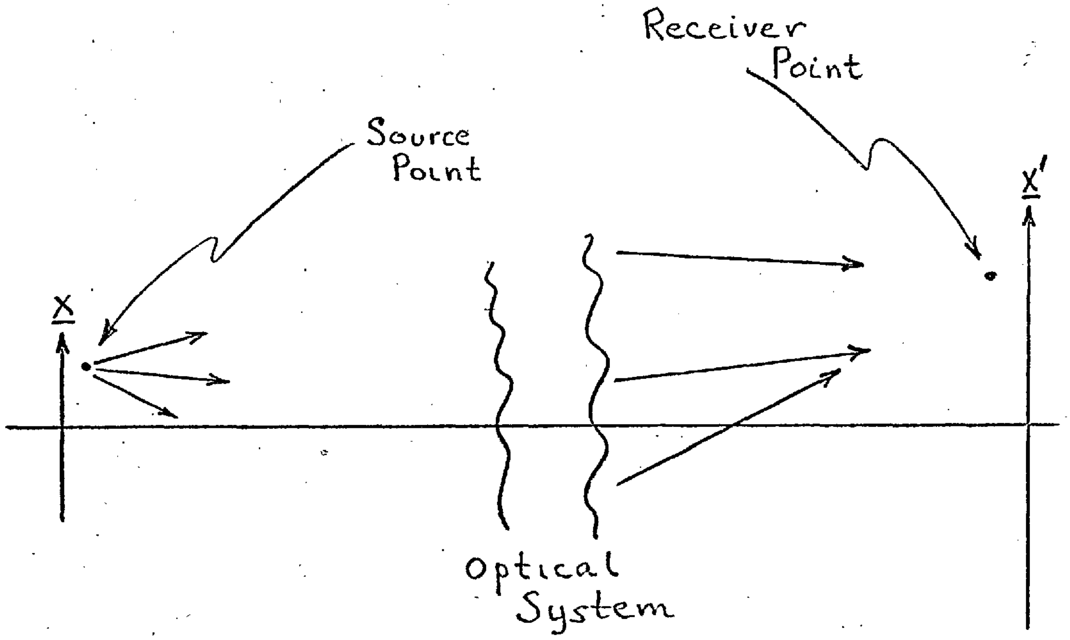


Figure 1

Source and Receiver Coordinates

physically correct, because that part of $S(t,\underline{x})$ in the source disturbance function which has not yet arrived at the detector at time $t = t'$ does not yet exist insofar as the detector is concerned. The disturbance producing the response of the detector at $t = t'$ is thus $S'(t,\underline{x})$ for $t \leq t'$, but is zero for $t > t'$. We therefore define modified Fourier components for the disturbance as "seen" by the detector, denoting this truncated disturbance by

$$S'_{t'}(t,\underline{x}') \begin{cases} = S'(t,\underline{x}') & \text{for } t \leq t' \\ = 0 & \text{for } t > t', \end{cases} \qquad (3)$$

with a running Fourier spectrum

$$s'_{t'}(\nu,\underline{x}') = \int\limits_{\substack{-\infty \\ t}}^{t'} S'(t,\underline{x}')\, e^{-i2\pi\nu t}\, dt = \int\limits_{\substack{-\infty \\ t}}^{\infty} S'_{t'}(t,\underline{x}')\, e^{-i2\pi\nu t}\, dt, \qquad (4)$$

and with

$$S'_{t'}(t,\underline{x}') = \int\limits_{\substack{-\infty \\ \nu}}^{\infty} s'_{t'}(\nu,\underline{x}')\, e^{i2\pi\nu t}\, d\nu .$$

Second, the standard Fourier representation implies the existence of sin/cos spectral components which exist for all time, for $t = \pm\infty$; even the modified (running) Fourier representation introduced above still implies the existence of these components for all times between $-\infty$ and t'. However, this is physically incorrect, and if these semi-infinite spectral components are used in an analysis of the fluctuations of $s_{t'}(v,\underline{x})$ [‡] as a function of t', wrong conclusions may ensue[*]. As it will shortly be necessary to consider such fluctuations, it is important to examine more carefully the relationship of the components of $s_{t'}(v,\underline{x})$ to the signal $S(t,\underline{x})$. Let $S(t,\underline{x})$ be the superposition of very many pulses, each pulse deriving from a single radiating atom, and each of roughly the length determined by the coherence time τ_c. By additivity, the spectral components of this superposition of pulses may themselves be considered a superposition of sin/cos terms, each term deriving from a particular pulse. These sin/cos terms are not physically infinitely extended in time, however, because for times when a particular pulse is zero, it is physically necessary that <u>each</u> spectral component deriving from that pulse (not only the sum of all components) also

[‡] (The typewritten symbol v is equivalent to the handwritten symbol ν.)

[*] To paraphrase Stone's (1963, p. 46), warning against the misuse of complex representation: "The use of Fourier theory in ways that do not give physically correct results is a pit which contains many victims."

be zero for such times. Were this not so, a mono-frequency detector could respond to a non-zero monochromatic spectral component even though, for that specific time, the signal itself were zero.[x] This means that as t' changes by roughly the coherence time τ_c, an entirely different set of signal pulses, and hence an entirely different set of sin/cos terms, must be considered. Although, with many pulses present, the resultant expected amplitude of $s_{t'}(v,\underline{x})$ will be independent of t', the phase of $s_{t'}(v,\underline{x})$ will be random between 0 and 2π for each different set of pulses, and hence the phase of $s_{t'}(v,\underline{x})$ will suffer large fluctuations at intervals on the order of the coherence time. (Davenport and Root, p. 161).

Let the total complex transmission for a given frequency v, from the source point $\underline{x}$ to the detector point $\underline{x}'$, be denoted by $u(v;\underline{x},\underline{x}')$. This means that a (fictional)

[x] It should be noted that this problem, the physical unreality of infinitely extended spectral components, does not usually cause difficulty in common problems of spectral analysis wherein a signal is known _a posteriori_ and only its spectrum is desired; however, in any problem involving the behaviour of a function in its reciprocal Fourier domain, as for example $s(v)$ in the t domain, the assumptions underlying the defining Fourier equations must be examined in the light of physical restrictions as well as mathematical convention.

monochromatic wavesource at $\underline{x}$, producing a wave of unit amplitude and zero phase, would produce at $\underline{x}'$ a disturbance of real amplitude equal to $|u|$ and phase equal to $\arg(u)$. This complex transmission will in general be frequency-dependent in both modulus and phase.

Analogous to $S'_{t'}(t,\underline{x}')$, the disturbance at the detector, denote the disturbance at a source point $\underline{x}$, at $t = t_o$, by $S_{t_0}(t,\underline{x})$, where

$$S_{t_0}(t,\underline{x}) \begin{array}{ll} = S(t,\underline{x}) & \text{for } t \leq t_0 \\ = 0 & \text{for } t > t_0 \, . \end{array}$$

The time t is earlier than t' by Δ, the transit time from $\underline{x}$ to $\underline{x}'$, which is simply equal to $\frac{1}{2\pi\nu}\arg(u)$. The disturbance $S_{t_0}(t,\underline{x})$ may be analysed into a spectrum

$$s_{t_0}(\nu,\underline{x}) = \int\limits_{\substack{-\infty \\ t}}^{t_0} S(t,\underline{x})\, e^{-i2\pi\nu t} dt = \int\limits_{\substack{-\infty \\ t}}^{\infty} S_{t_0}(t,\underline{x})\, e^{-i2\pi\nu t} dt, \quad (5)$$

with

$$S_{t_0}(t,\underline{x}) = \int\limits_{\substack{-\infty \\ \nu}}^{\infty} s_{t_0}(\nu,\underline{x})\, e^{i2\pi\nu t} d\nu \, . \quad (6)$$

A Fourier component $s_{t_0}(\nu,\underline{x})$ in the source will produce at the detector a Fourier component with complex amplitude given by

$$s'_{t'}(\nu, \underline{x}') = s_{t_0}(\nu, \underline{x})\ u(\nu; \underline{x}, \underline{x}') \quad . \tag{7}$$

If the media between the source and detector are non-dispersive, the transit time Δ will be independent of v, and the disturbance $S_{t_0}(v,\underline{x})$ synthesized in equation (6) may be regarded as a real disturbance which has been produced at the source at time $t = t_0$ and propagated to arrive at the detector at time $t = t'$. However, if as is frequently the case, there are glass or other dispersive components in the optical system between the source and the detector, Δ will depend on v, and the different Fourier components arriving at the detector at time $t = t'$ will then have left the source at different times $t = t_0$. In these cases, the disturbance synthesized in equation (6) will not actually have existed physically at the source in that precise form.

The total instantaneous disturbance at the detector at $t = t'$ will be given by

$$S'(t', \underline{x}') = \int_{\underline{x}} \int_{-\infty}^{\infty} s_{t_0}(\nu, \underline{x})\ u(\nu; \underline{x}, \underline{x}')\ e^{i2\pi\nu t'} d\nu\, d\underline{x}, \tag{8}$$

and the total instantaneous power falling on the point $\underline{x}'$ of the detector at time $t = t'$ will be given by the square of (8), that is by

$$P(t',\underline{x}') = \iint\limits_{\underline{x}\,\hat{\underline{x}}} \iint\limits_{-\infty\,\nu\,\hat{\nu}}^{\infty} \mathcal{A}^{*}_{t_0}(\nu,\underline{x})\, \mathcal{A}_{t_0}(\hat{\nu},\hat{\underline{x}})\, u^{*}(\nu;\underline{x},\underline{x}')\, u(\hat{\nu};\hat{\underline{x}},\underline{x}')\, e^{i2\pi(\hat{\nu}-\nu)t'} \, d\nu\, d\hat{\nu}\, d\underline{x}\, d\hat{\underline{x}} \quad (9),$$

the integrations over $\underline{x}$, $\hat{\underline{x}}$ being over the spatial extent of the source. The symbol $\wedge$ is used only to distinguish between the variables of the double integration.

For any detector, the signal produced by a beam of monochromatic light will depend on wavelength. Let $\Lambda(v)$ be this spectral sensitivity as a function of frequency. Each term in (9) is a cross-product of a signal of frequency v with one of frequency $\hat{v}$, and the question arises as to what spectral sensitivity factor should be used when two signals of different frequencies are combined. A similar problem exists when two highly coherent beams of slightly different frequencies are heterodyned, giving a beat frequency. In such a case, the photons in the beam will still be those of energies corresponding to the separate frequencies of the two coherent beams, and not of frequency corresponding to the beat frequency. Since each of these beams acting alone has an effective amplitude $\sqrt{\Lambda(\nu)}\,\mathcal{A}(\nu)$, it is reasonable to multiply the cross-product terms in (9) by $\sqrt{\Lambda(\nu)\,\Lambda(\hat{\nu})}$ to obtain the signal produced. This signal, for example a photocurrent, produced in the detector at any instant $t = t'$ is thus:

$$I(t', \underline{x}') = \iint\limits_{\underline{x}\,\hat{\underline{x}}} \iint\limits_{v\,\hat{v}\,-\infty}^{\infty} \sqrt{\Lambda(v)\Lambda(\hat{v})}\, \mathcal{A}^{*}_{t_\circ}(v, \underline{x})\, \mathcal{A}_{t_\circ}(\hat{v}, \hat{\underline{x}})\; \mathcal{U}^{*}(v; \underline{x}, \underline{x}')$$

$$\mathcal{U}(\hat{v}; \hat{\underline{x}}, \underline{x}')\; e^{i2\pi(\hat{v}-v)t'}\, dv\, d\hat{v}\, d\underline{x}\, d\hat{\underline{x}}\,. \qquad (10)$$

This formula may be interpreted as an integration over the beat frequencies $\mu = \hat{v}-v$ in the photocurrent. The detector circuit will have a frequency response which is significantly different from zero only in a finite bandwidth $\mu_\circ$, so that for a given v, the effective range of $\hat{v}$ will be from $\mu = v + \frac{\delta v}{2}$ to $\mu = v - \frac{\delta v}{2}$, with $\delta v = |\hat{v}-v|_{max} = \mu_\circ$. This maximum beat frequency $\mu_\circ$ corresponds to a maximum wavelength different $\delta\lambda_\circ$, where $\delta\lambda_\circ = \frac{\lambda^2}{c}\,\mu_\circ$. For $\lambda_\circ = 5 \times 10^{-5}$ cm, $\delta\lambda_\circ = 8\mu_\circ \times 10^{-12}$ Å. Thus, even for $\mu_\circ$ as great as 10^6, $\delta\lambda_\circ$ has only the value 8×10^{-6} Å. Over this wavelength range, and for laboratory-scale path lengths, both $\Lambda(v)$ and $\mathcal{U}(v)$ will be constant, and one may write

$$\Lambda(\hat{v}) = \Lambda(v+\mu) = \Lambda(v)$$

$$\mathcal{U}(\hat{v}; \hat{\underline{x}}, \underline{x}') = \mathcal{U}(v; \hat{\underline{x}}, \underline{x}') \qquad (11)$$

giving, with the substitution $\hat{v}-v = \mu$ and a change in the order of integration,

$$I(t',\underline{x}') = \int_{\mu}\Big[\int_{\nu}\int_{\underline{x}}\int_{\hat{\underline{x}}}\Lambda(\nu)\, U^{*}(\nu;\underline{x},\underline{x}')\, U(\nu;\hat{\underline{x}},\underline{x}')\, s^{*}_{t_0}(\nu,\underline{x})\, s_{t_0}(\nu+\mu,\hat{\underline{x}})\, d\underline{x}\, d\hat{\underline{x}}\, d\nu\Big]\, e^{i2\pi\mu t'}\, d\mu \qquad (12)$$

for the photocurrent detected by the circuit.

The complex amplitude at time t' of the beat frequency μ will be determined by the expression in the square brackets in (12), that is by

$$\int_{\nu}\int_{\underline{x}}\int_{\hat{\underline{x}}}\Lambda(\nu)\, U^{*}(\nu;\underline{x},\underline{x}')\, U(\nu;\hat{\underline{x}},\underline{x}')\, s^{*}_{t_0}(\nu,\underline{x})\, s_{t_0}(\nu+\mu,\hat{\underline{x}})\, d\underline{x}\, d\hat{\underline{x}}\, d\nu . \qquad (13)$$

Because the factors $s^{*}_{t_0}(v,\underline{x})$ and $s_{t_0}(v+\mu,\hat{\underline{x}})$ will each jump in phase at intervals of the order of the coherence time, the complex amplitude of the beat frequency μ will also fluctuate with time. Also, any correlation between the arguments of $s^{*}_{t_0}(v,\underline{x})$ and $s_{t_0}(v+\mu,\hat{\underline{x}})$ will tend to be smaller the more $\hat{\underline{x}}$ differs from $\underline{x}$, and thus the integral over $\hat{\underline{x}}$ will be restricted to a small area centered on $\underline{x}$. Finally, the phases of the contributions to a given beat frequency μ, arising from different frequency pairs, such as v_1, $v_1+\mu$ and v_2, $v_2+\mu$, will also be less correlated the greater the difference between v_1 and v_2. The effect of these factors in the integrations over $\underline{x}$, $\hat{\underline{x}}$, and v will be to give small and temporally fluctuating amplitudes to the non-zero beat frequencies μ . Thus the fluctuations in (13) will tend to average to a small noise on a steady signal. This steady signal will be given by the time average

$$\frac{1}{\tau} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} I(t', \underline{x}') \; dt' \qquad (14)$$

where τ is a time which is short compared with $1/\mu_o$, the resolving time of the detector circuit, but long enough to include a representative sample of the noise. This means that τ should be much greater than the coherence time.

It follows from these arguments that, to increase the detected noise fluctuations demands a spatially small source, a narrow spectral width, and a high bandwidth for the detector circuits. These are, of course, the conditions chosen for photon-bunching experiments, wherein the fluctuations rather than the d.c. signal are of interest.

The above considerations refer only to fluctuations in the power associated with beat-frequency terms, assuming the mean power for each frequency to remain constant over short intervals of time. They do not, therefore, include Poisson fluctuations in the photon emission; in the present treatment, the level of radiation density is assumed sufficiently high that the Poisson fluctuations are negligible.

The steady signal defined by (14) is given, using (12), by

$$I(\underline{x}') = \int_{\nu} \Lambda(\nu) \iint_{\underline{x}\ \hat{\underline{x}}} u^{*}(\nu;\underline{x},\underline{x}')\ u(\nu;\hat{\underline{x}},\underline{x}')$$
$$\left\{ \int_{\mu} \frac{1}{\tau} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} s^{*}_{t_0}(\nu,\underline{x})\ s_{t_0}(\nu+\mu,\hat{\underline{x}})\ e^{i2\pi\mu t'}\, dt'\, d\mu \right\} d\underline{x}\, d\hat{\underline{x}}\, d\nu . \tag{15}$$

The variations of $s^{*}_{t_0}(v,\underline{x})$ and $s_{t_0}(v+\mu,\hat{\underline{x}})$ occur on a time scale of the order of the coherence time τ_c . For the condition $1/\mu_0 \gg \tau_c$, that is for a detector circuit of resolving time much longer than the coherence time of the radiation, the exponential factor $e^{i2\pi\mu t}$ will change only very slowly compared to $s^{*}_{t_0}(v,\underline{x})\ s_{t_0}(v+\mu,\hat{\underline{x}})$, so that this latter product may be replaced by its time average, and then the expression in the curly brackets in (15) will give

$$\int_{\mu} \frac{1}{\tau} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} s^{*}_{t_0}(\nu,\underline{x})\ s_{t_0}(\nu+\mu,\hat{\underline{x}})\ e^{i2\pi\mu t'}\, dt'\, d\mu =$$

$$= \int_{\mu} \langle s^{*}(\nu,\underline{x})\ s(\nu+\mu,\hat{\underline{x}}) \rangle \frac{1}{\tau} \int_{-\frac{\tau}{2}}^{\frac{\tau}{2}} e^{i2\pi\mu t'}\, dt'\ d\mu . \tag{16}$$

For the effective range of μ , which corresponds in a typical case to $\delta\lambda \approx 10^{-5}$ Å , $s_{t_0}(v+\mu,\hat{\underline{x}})$, as determined by (5), may be replaced by $s_{t_o}(v,\hat{\underline{x}})$. In the time averaging operation denoted by the angular brackets $\langle\ \rangle$, the subscript t_o is omitted, it being assumed that the spectrum of the light from the source is stationary in a statistical sense. Integrating with respect to t , (16) becomes

$$\langle s^*(\nu,\underline{x})\, s(\nu,\hat{\underline{x}})\rangle \int_{-\infty}^{\infty} \frac{\sin \pi\mu\tau}{\pi\mu\tau}\, d\mu = \frac{1}{\tau}\langle s^*(\nu,\underline{x})\, s(\nu,\hat{\underline{x}})\rangle , \qquad (17)$$

and substitution of this value of (17) into (15) now gives

$$I(\underline{x}') = \int_{\nu} \Lambda(\nu) \iint_{\underline{x}\,\hat{\underline{x}}} \frac{1}{\tau}\langle s^*(\nu,\underline{x})\, s(\nu,\hat{\underline{x}})\rangle\, u^*(\nu;\underline{x},\underline{x}')\, u(\nu;\hat{\underline{x}},\underline{x}')\, d\underline{x}\, d\hat{\underline{x}}\, d\nu \qquad (18)$$

for the steady detected signal.

Consider now that $s(v,\underline{x})$, the Fourier spectrum of $S(t,\underline{x})$, is only proportional, and not equal, to the complex amplitude spectrum of $S(t,\underline{x})$, which will be denoted $\alpha(v,\underline{x})$. Equating the total mean power in the signal $S(t,\underline{x})$ and in its complex amplitude spectrum $\alpha(v,\underline{x})$, we have

$$\frac{1}{\tau}\int_{t} |S(t,\underline{x})|^2\, dt = \int_{\nu} |\alpha(\nu,\underline{x})|^2\, d\nu \qquad (19)$$

which gives, using Rayleigh's theorem on the left hand side,

$$\frac{1}{\tau}\int_{\nu} |\mathscr{S}(\nu,\underline{x})|^2\, d\nu = \int_{\nu} |\alpha(\nu,\underline{x})|^2\, d\nu \tag{20}$$

so that, for any given frequency

$$\frac{1}{\sqrt{\tau}}\, |\mathscr{S}(\nu,\underline{x})| = |\alpha(\nu,\underline{x})|\,. \tag{21}$$

This expression relates the modulus of the Fourier transform of $S(t,\underline{x})$ and the modulus of the physical complex amplitude associated with $S(t,\underline{x})$. Equation (15) can now be written

$$I(\underline{x}) = \int_{\nu}\Lambda(\nu)\iint_{\underline{x}\,\hat{\underline{x}}}\langle\alpha^{*}(\nu,\underline{x})\,\alpha(\nu,\hat{\underline{x}})\,\mathcal{U}^{*}(\nu;\underline{x},\underline{x}')\,\mathcal{U}(\nu;\hat{\underline{x}},\underline{x}')\, d\underline{x}\, d\hat{\underline{x}}\, d\nu. \tag{22}$$

The factor $\langle\alpha^{*}(\nu,\underline{x})\alpha(\nu,\hat{\underline{x}})\rangle$ provides a phenomenological basis for defining the micro-coherence between any two points $\underline{x}$, $\hat{\underline{x}}$ of a thermal source. If $E(v,\underline{x})$ denotes the mean power from $\underline{x}$ at frequency v , then

$$E(\nu,\underline{x}) = \langle\alpha^{*}(\nu,\underline{x})\,\alpha(\nu,\underline{x})\rangle = \langle|\alpha(\nu,\underline{x})|^{2}\rangle, \tag{23}$$

and equation (22) may be written

$$I(\underline{x}') = \int_{\nu}\Lambda(\nu)\iint_{\underline{x}\;\hat{\underline{x}}}\sqrt{E(\nu,\underline{x})\,E(\nu,\hat{\underline{x}})}\;\Gamma(\nu;\underline{x},\hat{\underline{x}})$$

$$\mathcal{U}^{*}(\nu;\underline{x},\underline{x}')\;\mathcal{U}(\nu;\hat{\underline{x}},\underline{x}')\;d\underline{x}\;d\hat{\underline{x}}\;d\nu \tag{24}$$

where the micro-coherence of the source is specified by the factor

$$\Gamma(\nu;\underline{x},\hat{\underline{x}}) = \frac{\langle \alpha^*(\nu,\underline{x})\ \alpha(\nu,\hat{\underline{x}})\rangle}{\sqrt{\langle|\alpha(\nu,\underline{x})|^2\rangle\langle|\alpha(\nu,\hat{\underline{x}})|^2\rangle}} \tag{25}$$

or,

$$\Gamma(\nu;\underline{x},\hat{\underline{x}}) = \frac{\langle \alpha^*(\nu,\underline{x})\ \alpha(\nu,\hat{\underline{x}})\rangle}{\sqrt{E(\nu,\underline{x})\ E(\nu,\hat{\underline{x}})}}\ , \tag{26}$$

and (24) may be written

$$\mathcal{I}(\underline{x}') = \int_{\nu} \Lambda(\nu)\, \mathcal{I}(\nu,\underline{x}')\, d\nu \tag{27}$$

with

$$\mathcal{I}(\nu,\underline{x}') = \iint_{\underline{x}\ \hat{\underline{x}}} \sqrt{E(\nu,\underline{x})\, E(\nu,\hat{\underline{x}})}\ \Gamma(\nu;\underline{x},\hat{\underline{x}})\, u^*(\nu;\underline{x},\underline{x}')\, u(\nu;\hat{\underline{x}},\underline{x}')\ d\underline{x}\, d\hat{\underline{x}}\ . \tag{28}$$

Equation (28) gives the power of the light of frequency v at the point $\underline{x}'$, and (27) shows the signal from the detector to be given by the weighted sum of the powers of the different frequencies in the source. That is, the different frequencies in the light behave incoherently, when the coherence time is much shorter than the integrating time of the detector. This will be the case for any thermal source and existing detection techniques, and also for laser light unless extremely short detection times are used.

The factor $\Gamma(v;\underline{x},\underline{\hat{x}})$ thus defined specifies the spatial coherence for light of frequency v between disturbances originating at the points $\underline{x}$, $\underline{\hat{x}}$ of the source. Consider the case for a coherent source, putting $\Gamma(v;\underline{x},\underline{\hat{x}}) = 1$. Equation (28) then becomes

$$I(v,\underline{x}') = \left| \int_{\underline{x}} \sqrt{E(v,\underline{x})} \;\; \mathcal{U}(v;\underline{x},\underline{x}') \, d\underline{x} \right|^2 , \tag{29}$$

so that the power is found by assuming all points of the source to emit coherent and cophasal waves of real amplitude equal to the square root of the mean power radiated by the source.

Consider now the form of (28) when the source is completely incoherent. Noting that by the definition (25), $\Gamma(v;\underline{x},\underline{x}) = \Gamma(v;\underline{\hat{x}},\underline{\hat{x}}) = 1$, so that the form to be expected for $\Gamma(v;\underline{x},\underline{\hat{x}})$ is

$$\Gamma(v;\underline{x},\underline{\hat{x}}) \begin{array}{ll} = 1 & \text{for } \underline{\hat{x}} = \underline{x} \\ = 0 & \text{for } \underline{\hat{x}} \neq \underline{x} . \end{array} \tag{30}$$

However, this form substituted into (28) merely gives $I(v;\underline{x}') = 0$; the reason for this may be understood as follows. In the analysis used here, $S(t,\underline{x})$ is the disturbance produced by a unit volume element of the source, and $S(t,\underline{x})\delta\underline{x}$ is the disturbance produced by a volume element $\delta\underline{x}$. It follows that

$$\sqrt{E(v,\underline{x})} = \sqrt{\langle|\alpha(v,\underline{x})|^2\rangle} = \sqrt{\frac{1}{\tau}\langle|\mathcal{A}(v,\underline{x})|^2\rangle} \qquad (31)$$

is the time-averaged real amplitude per unit volume of the source. If $E_o(v,\underline{x})$ is now defined to be the mean power per unit volume, equating expressions for the mean power emitted by a volume element $\delta\underline{x}$ of the source gives

$$\left[\sqrt{E(v,\underline{x})}\,\delta\underline{x}\right]^2 = E_o(v,\underline{x})\,\delta\underline{x}$$

so that

$$E(v,\underline{x})\,\delta\underline{x} = E_o(v,\underline{x}) . \qquad (32)$$

Thus the square of the real amplitude per unit volume, $E(v,\underline{x})$, is only proportional to the power per unit volume, $E_o(v,\underline{x})$; these two quantities are not equal, and their factor of proportionality is seen to be the volume element $\delta\underline{x}$. This factor of proportionality tends to zero as $\delta\underline{x} \to 0$, and this accounts for the null result when (30) is used in (28). Assume now an incoherent source comprising volume elements of size $\delta\underline{x}$, so that, in the integration with respect to $\hat{\underline{x}}$ in (28), $\Gamma(v;\underline{x},\hat{\underline{x}}) = 0$ except for the single element $\delta\hat{\underline{x}} = \delta\underline{x}$. Equation (28) may then be written

$$I(v,\underline{x}') = \int_{\underline{x}} E(v,\underline{x})\,\delta\underline{x}\;\mathcal{U}^*(v;\underline{x},\underline{x}')\,\mathcal{U}(v;\underline{x},\underline{x}')\,d\underline{x} , \qquad (33)$$

or, by (32),

$$I(\nu,\underline{x}') = \int_{\underline{x}} E_o(\nu,\underline{x}) \, |\mathcal{U}(\nu;\underline{x},\underline{x}')|^2 \, d\underline{x} \ , \qquad (34)$$

where now the power per unit volume is employed. Thus, $\Gamma(\mathrm{v};\underline{x},\hat{\underline{x}})$ being the unit-function (30) corresponds to a completely incoherent source, the resultant power at $\underline{x}'$ being given by the sum of the separate powers produced there by the different elements of the source.

An alternative procedure may be used for the case of an incoherent source. The product

$$\Gamma(\nu;\underline{x},\hat{\underline{x}})\sqrt{E(\nu,\underline{x})\ E(\nu,\hat{\underline{x}})} = \frac{\Gamma(\nu;\underline{x},\hat{\underline{x}})}{\sqrt{\delta\underline{x}\ \delta\hat{\underline{x}}}}\sqrt{E_o(\nu,\underline{x})\,E_o(\nu,\hat{\underline{x}})}$$

tends, as $\delta\underline{x} \to 0$ and $\delta\hat{\underline{x}} \to 0$, to the form $\Gamma(\mathrm{v};\underline{x},\hat{\underline{x}}) = \infty$ for $\hat{\underline{x}} = \underline{x}$, and $\Gamma(\mathrm{v};\underline{x},\hat{\underline{x}}) = 0$ for $\hat{\underline{x}} \neq \underline{x}$; this suggests that a delta function may be used. Thus, one may put

$$\Gamma(\nu;\underline{x},\hat{\underline{x}})\sqrt{E(\nu,\underline{x})\ E(\nu,\hat{\underline{x}})} = \delta(\hat{\underline{x}}-\underline{x})\sqrt{E_o(\nu,\underline{x})E_o(\nu,\hat{\underline{x}})} \qquad (35)$$

and so, formally, one may write

$$\Gamma(\nu;\underline{x},\hat{\underline{x}}) = \delta(\hat{\underline{x}}-\underline{x})$$

to describe an ideally incoherent source. Equation (28) will then give

$$I(\nu,\underline{x}') = \int_{\underline{x}}\int_{\hat{\underline{x}}} \delta(\hat{\underline{x}}-\underline{x})\sqrt{E(\nu,\underline{x})\,E(\nu,\hat{\underline{x}})}\ \mathcal{U}^*(\nu;\underline{x},\underline{x}')\ \mathcal{U}(\nu;\hat{\underline{x}},\underline{x}') \ d\underline{x}\, d\hat{\underline{x}} \ ,$$

that is,

$$I(\nu,\underline{x}') = \int_{\underline{x}} E(\nu,\underline{x}) \left| U(\nu;\underline{x},\underline{x}') \right|^2 d\underline{x} . \qquad (36)$$

It has only to be remembered that for a perfectly incoherent source, $E(\nu,\underline{x})$ must be interpreted as the power per unit volume at the source point $\underline{x}$, as seen in (34) and (35), and not as the square of the real amplitude per unit volume.

The above difficulty has been noted previously, in a different form, by Beran and Parrent (1964, p. 57). These authors, in discussing an incoherent source in terms of an enclosing surface, conclude that for the spatial coherence of the source to be represented by a delta function, the intensity over the source would have to be infinite. They therefore conclude on mathematical grounds that a perfectly incoherent source $[\Gamma(\nu;\underline{x},\hat{\underline{x}}) = \delta(\hat{\underline{x}}-\underline{x})]$ is impossible. The treatment given here shows that the difficulty resides in the reduction of the two-dimensional integral (28) to the one-dimensional integral (33), and in going from the field disturbance per unit volume to the power per unit volume, rather than in whether the source coherence function is taken to the actual limiting form of the delta function. As shown above, a completely incoherent source causes no mathematical inconsistencies, and indeed one might postulate an extremely low pressure gas discharge as a realistic model for such a physical source.

To see in more detail the general significance of $\Gamma(v\,;\underline{x},\hat{\underline{x}})$, consider a source comprising only two elements $\delta\underline{x}_1 = \delta\underline{x}_2$, at $\underline{x}_1$ and $\underline{x}_2$ respectively. This is exactly analogous to the standard problem of computing the power distribution in a Young's fringe experiment, wherein the screen with the two holes is the equivalent to the present two-point source. The double integral (28) now reduces to four terms

$$\begin{aligned} I(\nu,\underline{x}') = {} & \Gamma(\nu;\underline{x}_1,\hat{\underline{x}}_1)\; E(\nu,\underline{x}_1)\,|\mathcal{U}(\nu;\underline{x}_1,\underline{x}')|^2(\delta\underline{x})^2 \\ & + \Gamma(\nu;\underline{x}_2,\underline{x}_2)\,E(\nu,\underline{x}_2)\,|\mathcal{U}(\nu;\underline{x}_2,\underline{x}')|^2(\delta\underline{x})^2 \\ & + \Gamma(\nu;\underline{x}_1,\underline{x}_2)\sqrt{E(\nu,\underline{x}_1)\,E(\nu,\underline{x}_2)}\,(\delta\underline{x})^2\,\mathcal{U}^*(\nu;\underline{x}_1,\underline{x}')\,\mathcal{U}(\nu;\underline{x}_2,\underline{x}') \\ & + \Gamma(\nu;\underline{x}_2,\underline{x}_1)\sqrt{E(\nu,\underline{x}_2)\,E(\nu,\underline{x}_1)}\,(\delta\underline{x})^2\,\mathcal{U}^*(\nu;\underline{x}_2,\underline{x}')\,\mathcal{U}(\nu;\underline{x}_1,\underline{x}'). \end{aligned} \tag{37}$$

From the definition (26),

$$\Gamma(\nu;\underline{x}_1,\underline{x}_1) = \Gamma(\nu;\underline{x}_2,\underline{x}_2) = 1$$

$$\Gamma(\nu;\underline{x}_1,\underline{x}_2) = \Gamma^*(\nu;\underline{x}_2,\underline{x}_1)\,.$$

Then, writing from (32)

$$\begin{aligned} E(\nu,\underline{x}_1)(\delta\underline{x})^2 &= E_0(\nu,\underline{x}_1)\delta\underline{x} = I_1(\nu) \\ E(\nu,\underline{x}_2)(\delta\underline{x})^2 &= E_0(\nu,\underline{x}_2)\delta\underline{x} = I_2(\nu) \end{aligned} \tag{38}$$

for the total powers radiated by the source elements at $\underline{x}_1$ and $\underline{x}_2$, (37) becomes

$$I(\nu,\underline{x}') = I_1(\nu)\,|u(\nu;\underline{x}_1,\underline{x}')|^2 + I_2(\nu)\,|u(\nu;\underline{x}_2,\underline{x}')|^2$$

$$+2\sqrt{I_1(\nu)I_2(\nu)}\,\mathrm{Re}\left\{\Gamma_{21}(\nu;\underline{x}_1,\underline{x}_2)\,u^*(\nu;\underline{x}_1,\underline{x}')\,u(\nu;\underline{x}_2,\underline{x}')\right\}. \qquad (39)$$

If now the distances from the two source points $\underline{x}_1$, $\underline{x}_2$ to an observation point $\underline{x}'$ are denoted by R_1, R_2, the complex transmission factors may be written

$$u(\nu;\underline{x}_1,\underline{x}') = \frac{1}{R_1}e^{-\frac{i2\pi c}{\nu}R_1} \quad \text{and} \quad u(\nu;\underline{x}_2,\underline{x}') = \frac{1}{R_2}e^{-\frac{i2\pi c}{\nu}R_2},$$

the partial powers may be normalized as

$$I_1'(\nu) = \frac{I_1(\nu)}{R_1^2} \quad \text{and} \quad I_2'(\nu) = \frac{I_2(\nu)}{R_2^2},$$

and (39) becomes

$$I(\nu,\underline{x}') = I_1'(\nu) + I_2'(\nu) + 2\sqrt{I_1'(\nu)I_2'(\nu)}\,|\Gamma_{21}(\nu)|$$

$$\cos\left\{\frac{2\pi c}{\nu}(R_1-R_2) - \arg[\Gamma_{21}(\nu)]\right\}.$$

This is the usual expression describing the Young's fringe experiment, with $|\Gamma_{21}(\nu)|$ giving the modulation of the fringes and $\arg[\Gamma_{21}(\nu)]$ giving their positioning. It may be remarked that the need to use the relations (38) to obtain this standard result correctly, re-emphasizes the distinction between the square of the amplitude per unit volume in the source and the power per unit volume.

A useful general conclusion regarding thermal light sources may be drawn from (27) and (28). In these formulae, $\mathcal{U}(\nu;\underline{x},\underline{x}')$ is the complex amplitude produced at $\underline{x}'$ by a (fictitious) monochromatic wave of frequency v with unit amplitude and zero phase, at the point $\underline{x}$ in the source. Thus the product $\sqrt{E(\nu,\underline{x})}\,\mathcal{U}(\nu;\underline{x},\underline{x}')$ will be equal to the complex amplitude which would be produced at $\underline{x}'$ by a monochromatic wave of real amplitude $\sqrt{E(\nu,\underline{x})}$ and zero phase emitted by the source element at $\underline{x}$. If the real amplitude $\sqrt{E(\nu,\underline{x})}$ is absorbed in the factor $\mathcal{U}(\nu;\underline{x},\underline{x}')$, so that

$$\mathcal{U}'(\nu;\underline{x},\underline{x}') = \sqrt{E(\nu;\underline{x})}\;\;\mathcal{U}(\nu;\underline{x},\underline{x}') , \tag{40}$$

(28) will then become

$$I(\nu,\underline{x}') = \iint\limits_{\underline{x}\;\hat{\underline{x}}} \Gamma(\nu;\underline{x},\hat{\underline{x}})\;\mathcal{U}'^{*}(\nu;\underline{x},\underline{x}')\;\mathcal{U}'(\nu;\hat{\underline{x}},\underline{x}')\,d\underline{x}\,d\hat{\underline{x}} . \tag{41}$$

This formula, in conjunction with (27), demonstrates that it is legitimate to say that the resultant intensity produced at any point by a thermal light source may be regarded as arising from perfectly monochromatic waves of real amplitude equal to the square root of the intensity in the source for each frequency, and between which there is a complex degree of spatial coherence for each frequency. This, in addition to the incoherence between different frequencies, shown in the discussion preceding equation (17), provides the following rules for the composition of light disturbances for any source wherein the coherence time is much shorter than the integrating time of the detector:

i) Fourier spectral components of the same frequency coming from the same element of the source may be considered to be perfectly coherent.

ii) Fourier spectral components of the same frequency coming from different elements of the source may be considered to be partially coherent, the degree of partial coherence being determined by the properties of the source.

iii) Fourier spectral components of different frequencies may be considered to be completely incoherent.

In Hopkins' treatment of partial coherence for incoherent sources, each point of an incoherent source is simply assumed to radiate independently of the other points, and rule (ii) is accordingly modified to state this.

Rule (ii) might seem problematical, because infinitely extended pure frequency components are postulated which are only partially coherent, yet by the ordinary meaning of coherence as the measure of the phase correlation between two signals, such components would have to be *ipso facto* perfectly coherent. However, it must be remembered that this is a *rule*, for use in solution of appropriate problems, and not a description of physical reality. Although infinitely extended Fourier components do not physically exist, the treatment given above shows that it is permissible to postulate that they do, and then to postulate additionally an *effective* degree of partial correlation of phase between these imagined pure frequency components. These three rules thus not only apply usefully to practical instrumental problems, but are physically correct in spite of seeming inconsistency in a strict mathematical sense. It is important to bear in mind, however, that these rules apply only where the coherence time of the radiation is much shorter than the integrating time of the detector.

Coherence at a Plane Illuminated by a Micro-Coherent Source

In addition to determining the power distribution at a plane illuminated by a micro-coherent source, using (28), it is also desirable to be able to calculate the spatial coherence at such a plane. With reference to Fig. 2, let $f_1(v,\underline{x}')$, $f_2(v,\underline{x}')$ be the complex transmission factors between points 1, 2 respectively, and $\underline{x}'$. From equation (40), $U'(\nu;\underline{x},\underline{x}')$ is just the total complex amplitude of frequency v from $\underline{x}$ at $\underline{x}'$, which is simply the sum of the complex amplitudes arriving from points 1 and 2. If

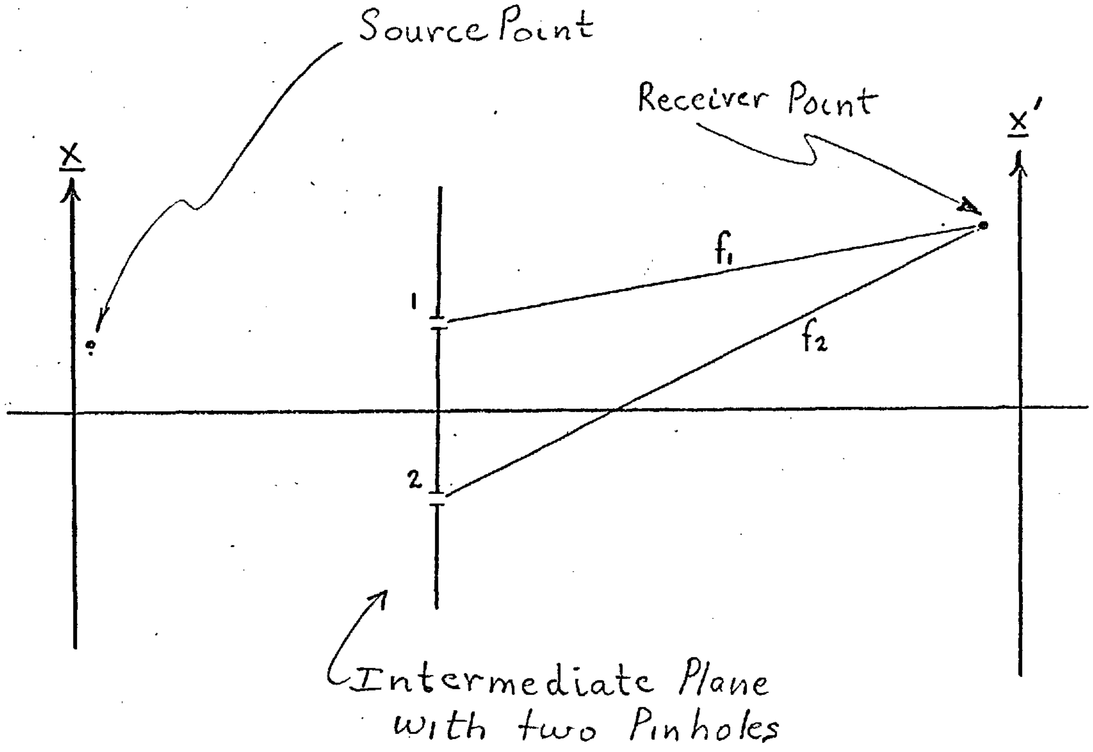


Figure 2

Source Illuminating an Intermediate Plane with Two Pinholes, and Subsequent Optical Paths f_1 , f_2 to a Receiver Point.

the complex amplitudes at points 1 and 2 deriving from $\underline{x}$ only are $U_1(\nu, \underline{x})$ and $U_2(\nu, \underline{x})$, we have

$$U'(\nu; \underline{x}, \underline{x}') = U_1(\nu, \underline{x}) f_1(\nu, \underline{x}') + U_2(\nu, \underline{x}) f_2(\nu, \underline{x}') . \qquad (42)$$

Consider the factor $U'^{*}(\nu; \underline{x}, \underline{x}') U(\nu; \hat{\underline{x}}, \underline{x}')$ in equation (41):

$$\begin{aligned} U'^{*}(\nu; \underline{x}, \underline{x}') U(\nu; \hat{\underline{x}}, \underline{x}') &= \\ &= [U_1^{*}(\nu, \underline{x}) f_1^{*}(\nu, \underline{x}') + U_2^{*}(\nu, \underline{x}) f_2^{*}(\nu, \underline{x}')][U_1(\nu, \hat{\underline{x}}) f_1(\nu, \underline{x}') + U_2(\nu, \hat{\underline{x}}) f_2(\nu; \underline{x}')] = \\ &= U_1^{*}(\nu, \underline{x}) U_1(\nu, \hat{\underline{x}}) |f_1(\nu, \underline{x}')|^2 + U_2^{*}(\nu, \underline{x}) U_2(\nu, \hat{\underline{x}}) |f_2(\nu, \underline{x}')|^2 \qquad (43) \\ &\quad + U_1^{*}(\nu, \underline{x}) U_2(\nu, \hat{\underline{x}}) f_1^{*}(\nu, \underline{x}') f_2(\nu, \underline{x}') + U_1(\nu, \underline{x}) U_2^{*}(\nu, \underline{x}) f_1(\nu, \underline{x}') f_2^{*}(\nu, \underline{x}') . \end{aligned}$$

When these four terms are substituted into equation (41), four integrals result. The first two integrals include the signal at points 1, 2 from the entire source, denoted by

$$I_1(\nu) = \iint\limits_{\underline{x}\,\hat{\underline{x}}} \Gamma(\nu; \underline{x}, \hat{\underline{x}}) \; U_1^{*}(\nu, \underline{x}) U_1(\nu, \hat{\underline{x}}) \; d\underline{x}\, d\hat{\underline{x}} \qquad (44) .$$

$$I_2(\nu) = \iint\limits_{\underline{x}\,\hat{\underline{x}}} \Gamma(\nu; \underline{x}, \hat{\underline{x}}) \; U_2^{*}(\nu, \underline{x}) U_2(\nu, \hat{\underline{x}}) \; d\underline{x}\, d\hat{\underline{x}} .$$

The third and fourth integrals are complex conjugates, in addition to an unimportant reversal in order between $\underline{x}$, $\hat{\underline{x}}$. Using the relation $z + z^{*} = 2 \mathrm{Re}\, z$, we have

$$I(\underline{x}') = \int_{-\infty \atop \nu}^{\infty} \Lambda(\nu) \Big\{ I_1(\nu) \left| f_1(\nu, \underline{x}') \right|^2 + I_2(\nu) \left| f_2(\nu, \underline{x}') \right|^2 + \\ + \iint_{\underline{x}\,\underline{\hat{x}}} \Gamma(\nu; \underline{x}, \underline{\hat{x}})\, 2\, \mathrm{Re} \left[U_1^*(\nu, \underline{x})\, U_2(\nu, \underline{\hat{x}})\, f_1^*(\nu, \underline{x}')\, f_2(\nu, \underline{x}') \right] \\ d\underline{x}\, d\underline{\hat{x}} \Big\}\, d\nu. \tag{45}$$

which may be written as

$$I(\underline{x}') = \int_{-\infty \atop \nu}^{\infty} \Lambda(\nu) \Big\{ I_1(\nu) \left| f_1(\nu, \underline{x}') \right|^2 + I_2(\nu) \left| f_2(\nu, \underline{x}') \right|^2 + \\ + \sqrt{I_1(\nu)\, I_2(\nu)}\; 2\; \mathrm{Re} \left[\Gamma_{21}(\nu)\, f_1^*(\nu, \underline{x}')\; f_2(\nu, \underline{x}') \right] \Big\}\, d\nu \tag{46}$$

with

$$\Gamma_{21}(\nu) = \frac{1}{\sqrt{I_1(\nu)\;\; I_2(\nu)}} \iint_{\underline{x}\;\underline{\hat{x}}} \Gamma(\nu; \underline{x}, \underline{\hat{x}})\, U_1^*(\nu, \underline{x})\, U_2(\nu, \underline{\hat{x}})\, d\underline{x}\, d\underline{\hat{x}} \tag{47}$$

giving $\Gamma_{21}(\nu)$ as desired.

For an incoherent source, $\Gamma^{INCOH}(\nu; \underline{x}, \underline{\hat{x}}) = \delta(\underline{x} - \underline{\hat{x}})$, and using this in (47) gives

$$\Gamma_{21}^{INCOH}(\nu) = \frac{1}{\sqrt{I_1(\nu) I_2(\nu)}} \iint_{\underline{x}\;\underline{\hat{x}}} \delta(\underline{x} - \underline{\hat{x}})\;\; U_1^*(\nu, \underline{x})\, U_2(\nu, \underline{\hat{x}})\; d\underline{x}\, d\underline{\hat{x}} =$$

$$\Gamma_{21}^{INCOH}(\nu) = \frac{1}{\sqrt{I_1(\nu) I_2(\nu)}} \int_{\underline{x}} U_1^*(\nu, \underline{x})\;\; U_2(\nu, \underline{x})\;\; d\underline{x} \tag{48}$$

which is the usual formula for the degree of spatial coherence between two points illuminated by an incoherent source.

For a coherent source $\Gamma^{COH}(\nu; \underline{x}, \hat{\underline{x}}) = e^{i\phi(\underline{x}-\hat{\underline{x}})}$, and again using this in (47) gives

$$\Gamma_{21}^{COH}(\nu) = \frac{1}{\sqrt{I_1(\nu)\, I_2(\nu)}} \iint_{\underline{x}\,\hat{\underline{x}}} e^{i\phi(\underline{x}-\hat{\underline{x}})}\, u_1^*(\nu, \underline{x})\, u_2(\nu, \hat{\underline{x}})\, d\underline{x}\, d\hat{\underline{x}} \quad (49)$$

whose squared modulus is

$$\left|\Gamma_{21}^{COH}(\nu)\right|^2 = \frac{1}{I_1(\nu)\, I_2(\nu)} \iint_{\underline{x}\,\hat{\underline{x}}} \iint_{\underline{x}'\,\hat{\underline{x}}'} u_1^*(\nu, \underline{x})\, u_2(\nu, \hat{\underline{x}})\, u_1(\nu, \underline{x}')\, u_2^*(\nu, \hat{\underline{x}}')\, e^{i\phi(\underline{x}-\hat{\underline{x}})\,\phi(\hat{\underline{x}}'-\underline{x}')}\, d\underline{x}\, d\hat{\underline{x}}\, d\underline{x}'\, d\hat{\underline{x}}'. \quad (50)$$

From (44)

$$I_1(\nu) = \iint_{\underline{x}\,\hat{\underline{x}}} e^{i\phi(\underline{x}-\hat{\underline{x}})}\, u_1^*(\nu, \underline{x})\, u_1(\nu, \hat{\underline{x}})\, d\underline{x}\, d\hat{\underline{x}} \quad (51)$$

$$I_2(\nu) = \iint_{\underline{x}'\,\hat{\underline{x}}'} e^{i\phi(\underline{x}'-\hat{\underline{x}}')}\, u_2^*(\nu, \underline{x}')\, u_2(\nu, \hat{\underline{x}}')\, d\underline{x}'\, d\hat{\underline{x}}', \quad (52)$$

and thus if the variables of integration $\underline{x}'$ and $\hat{\underline{x}}'$ are interchanged,

$$I_1(\nu)\, I_2(\nu) = \iint_{\underline{x}\,\hat{\underline{x}}} \iint_{\hat{\underline{x}}'\,\underline{x}'} e^{i\phi(\underline{x}-\hat{\underline{x}})\,\phi(\hat{\underline{x}}'-\underline{x}')}\, u_1^*(\nu, \underline{x})\, u_2^*(\nu, \hat{\underline{x}}')\, u_1(\nu, \hat{\underline{x}})\, u_2(\nu, \underline{x}')\, d\underline{x}\, d\hat{\underline{x}}\, d\hat{\underline{x}}'\, d\underline{x}' \quad (53)$$

which is identical to the integral factor in (50), thus giving $\left| \Gamma_{21}^{COH}(\nu) \right|^2 = 1 = \left| \Gamma_{21}^{COH}(\nu) \right|$.

Thus, (47) reduces to the correct expressions for each of the extreme cases of complete coherence and incoherence.

Coherence over One Resolution Element at a Source

If the micro-coherent source is viewed with an optical system having a resolution element of radius ρ, (Fig. 3), (27) may be simplified by separating the double integration over the source coordinates $\underline{x}$, $\hat{\underline{x}}$ into two stages of integration, the inner integral over an area the size of a resolution element, and the outer integral over all of the resolution elements comprising the source. Since the path differences between all points within a resolution element and all points in the viewing entrance pupil are equal to within $< \lambda/4$, as $\underline{x}$, $\hat{\underline{x}}$ vary over a given resolution element, $\arg[U'(\nu; \underline{x}, \underline{x}')]$ does not vary more than $\pi/2$ in phase. Additionally, $|U'(\nu; \underline{x}, \underline{x}')|$ is essentially constant for a geometrically uniform light source. Therefore to a close approximation we may say that $U'(\nu; \underline{x}, \underline{x}')$ is independent of $\underline{x}$ within the small area of a resolution element, and therefore may be removed from the inner integral over that resolution element. If we let $\underline{x}_0$ be the vector coordinate of the centre of a resolution element, we have from (28)

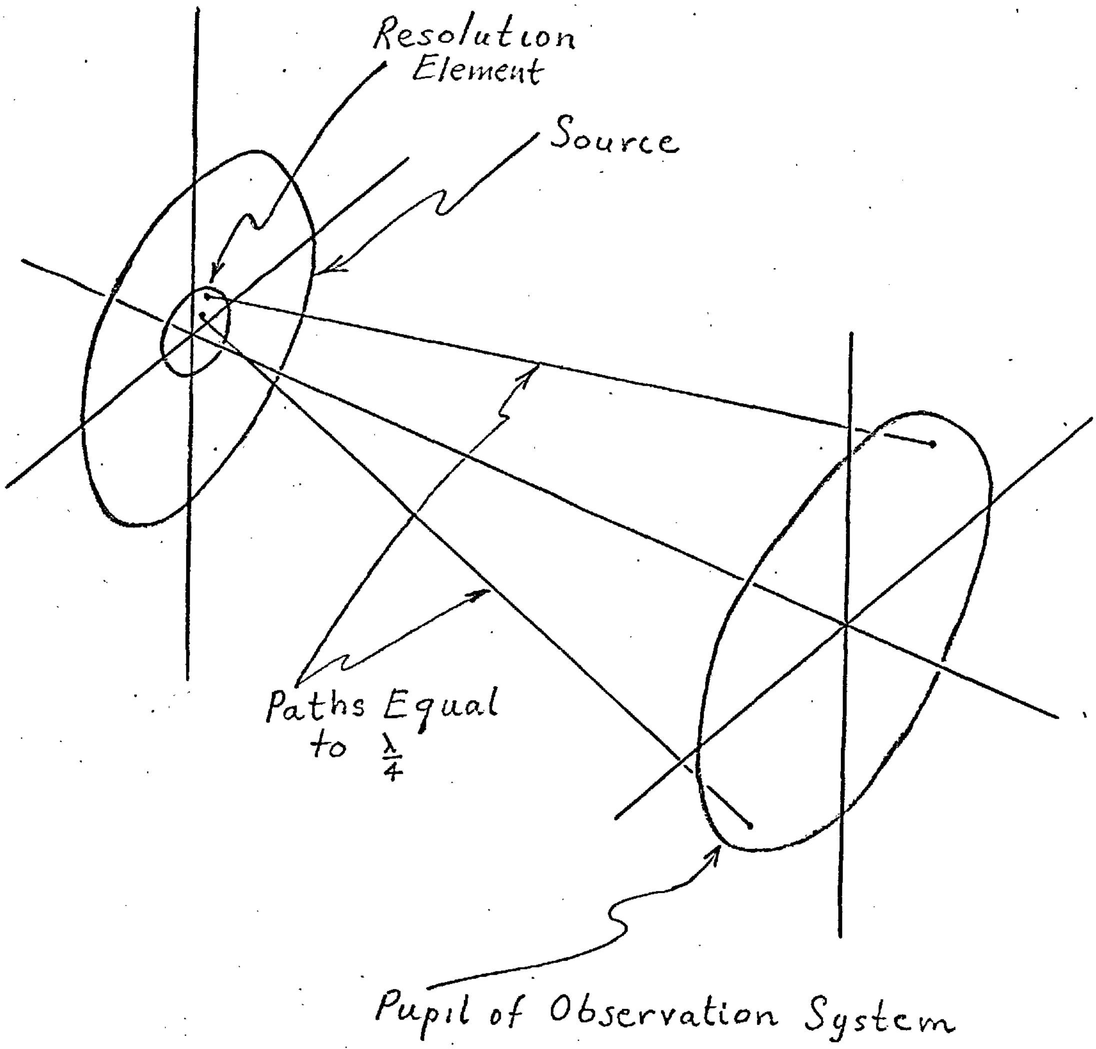

Resolution Element
Source
Paths Equal to $\frac{\lambda}{4}$
Pupil of Observation System

Figure 3

A Small Resolution Element, over which All Paths through the Observing Entrance Pupil are Equal to within $\frac{\lambda}{4}$, on a Large Source.

$$I(\underline{x}') = \int_{-\infty}^{\infty} \Lambda(\nu) \iint_{\underline{x}_0\, \hat{\underline{x}}_0} U'^{*}(\nu; \underline{x}, \underline{x}')\, U'(\nu; \hat{\underline{x}}, \underline{x}') \iint_{\rho} \Gamma(\nu; \underline{x}, \hat{\underline{x}})\, d\rho \quad d\underline{x}_0\, d\hat{\underline{x}}_0 \,. \qquad (54)$$

The inner integral, over a single resolution element ρ, is simply the average value of $\Gamma(\nu; \underline{x}, \hat{\underline{x}})$ over an area the size of a resolution element, and does not depend explicitly on the form of $\Gamma(\nu; \underline{x}, \hat{\underline{x}})$ over that area. This is only an analytic description of what we should reasonably expect, namely that because the viewing system cannot distinguish points inside an area the size of its resolution element, it can only average the radiation coming from such an area.

Vector Considerations of Fields Arising from a Collection of Dipoles

It has been assumed above that the disturbance $S(t,\underline{x})$ can properly be represented by a scalar, and does not require a vector representation. This is usually justifiable if the disturbance is in the far-field of the (assumed) radiating dipoles in the source, and is viewed over a small solid angle.

Imagine using a simple Young's two-pinhole screen to measure the coherence $\Gamma_{21}(\nu)$ at a plane illuminated by a source, in order to deduce information about the source micro-coherence function $\Gamma(\nu; \underline{x}, \hat{\underline{x}})$. If the light source is not spatially very small, then $\Gamma_{21}(\nu)$ will be appreciable only over very small distances, as can be seen from the standard Zernike-vanCittert theorem. In this case the two pinholes must be kept close together in order to obtain reasonable contrast in the fringes and hence precision in the final measurement of $\Gamma(\nu; \underset{\sim}{x}, \hat{\underset{\sim}{x}})$, and scalar theory is quite adequate. However, the Young's pinhole system would then have very low angular resolution, and could not determine significant information about the coherence function $\Gamma(\nu; \underset{\sim}{x}, \hat{\underset{\sim}{x}})$ over small distances on the source. It is therefore clear that we would have to use a small source and Young's pinholes widely spaced. However, in this case, any given dipole radiator in the source will then be under examination over a wide angle, and it is not obvious that scalar theory is still justifiable. For instance, if the source were to consist of a single dipole radiator, the measurement of $\Gamma_{21}(\nu)$ will be strongly affected by the angle subtended by the two pinholes and the orientation of the pinholes relative to the dipole, as can be seen with reference to Fig. 4. It is clear that $\Gamma_{21}(\nu) = 1$ as expected from a "point source," but that $\Gamma_{23}(\nu) = 0$, since in (47) both $I_3(\nu)$ and $U_3(\nu) = 0$. Thus, for wide separations of the pinholes, dipoles in at least some orientations would degrade the measurement of

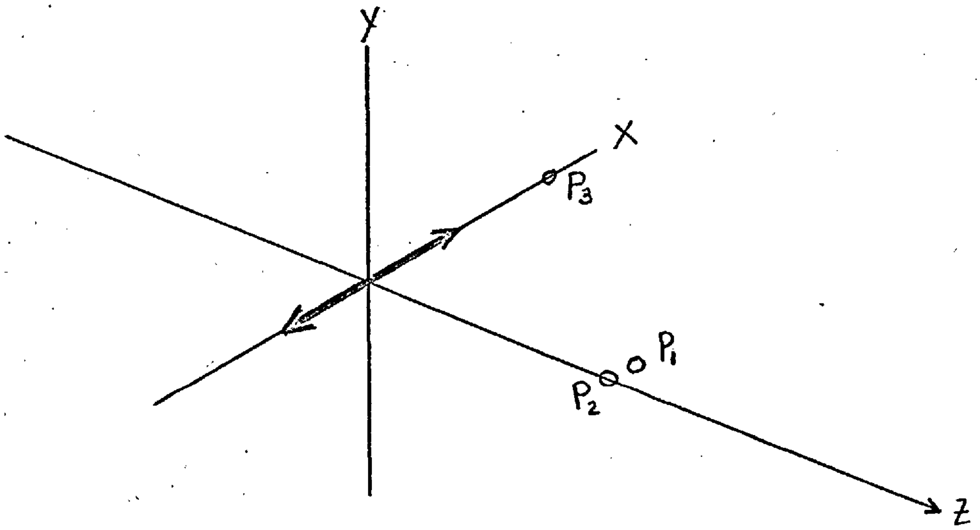


Figure 4

Three Observation Points P_1 , P_2 , P_3 ,

about a Dipole Radiator

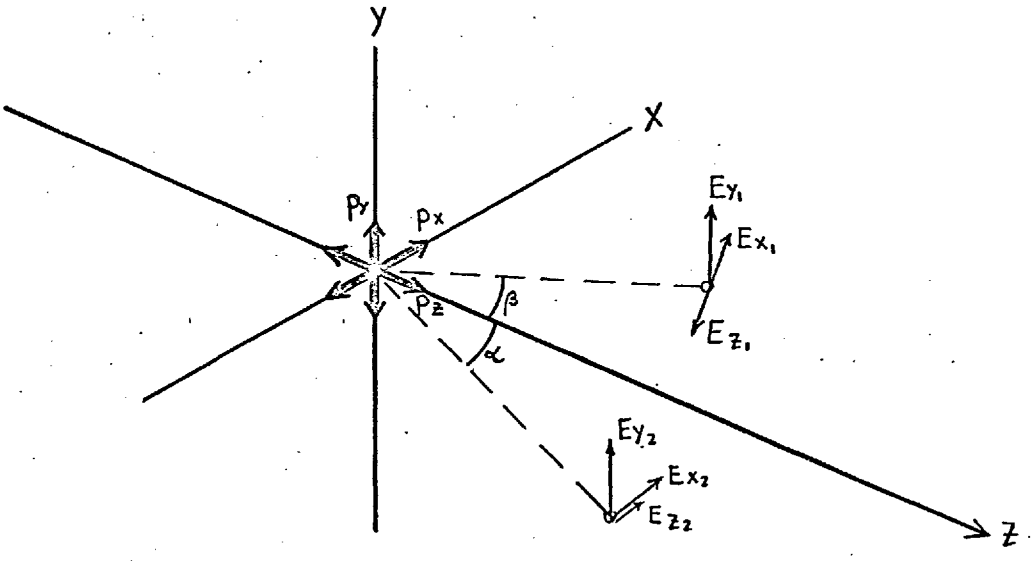


Figure 5

Dependence of Field Components, about the Three Unit Dipole Radiators, with Observation Position.

$\Gamma(\nu; \underset{\sim}{x}, \hat{\underset{\sim}{x}})$ by contributing only constant background power to the Young's fringe pattern.

However, if only the polarization component perpendicular to the plane containing the dipole and the two pinholes is used, this degradation does not occur. This can be seen with reference to Fig. 5, in which are shown three unit dipoles p_x, p_y, p_z aligned along the coordinate axes, and two pinholes in the $X-Z$ plane making arbitrary angles α , β to the Z axis and at equal distances from the origin. The magnitudes of the sums of the X and Z field components at the two pinholes, $E_{x_1} + E_{z_1}$ and $E_{x_2} + E_{z_2}$, will depend on α and β , and will in general not be equal. However, the y components, E_{y_1} and E_{y_2} , are equal for both pinholes. Any randomly oriented dipole may, in the far field, by analysed into a linear vector super-position of three axial dipoles, and if only the y polarization component is selected, each dipole will contribute fully to the contrast in the Young's fringe pattern, and scalar theory may be applied to this y component.

Isotropy of Source Micro-Coherence

To consider the question of the expected isotropy of source micro-coherence, examine the coherence relationships between three points in a micro-coherent source. Let these three points be at vector coordinates $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$, further lying at corners of an equilateral triangle, as shown in Fig. 6. The three points may be grouped in three pairs, each having an appropriate coherence relationship, denoted Γ_{12} , Γ_{13} , and Γ_{23} . These coherence functions may be written in terms of their moduli and arguments as $\Gamma_{12} = V_{12}\, e^{i\beta_{12}}$, $\Gamma_{13} = V_{13}\, e^{i\beta_{13}}$, and $\Gamma_{23} = V_{23}\, e^{i\beta_{23}}$, where V may be regarded as giving the degree of phase correlation between a pair of points and β as giving the phase delay.

In a thermally radiating source, each small volume of the source should radiate isotropically into the sphere surrounding it, and the coherence between two points of the source volume should depend only on the distance between them and not on the direction of their separation. This means that $\Gamma_{12} = \Gamma_{13}$, because $|\underline{x}_1 - \underline{x}_2| = |\underline{x}_1 - \underline{x}_3|$. It also means that $\Gamma_{12} = \Gamma_{13} = \Gamma_{23}$, for similar reasons. From this follows that $\beta_{12} = \beta_{13} = \beta_{23}$, which can be satisfied only for $\beta_{12} = \beta_{13} = \beta_{23} \equiv 0$; thus, the micro-coherence function Γ must be real.

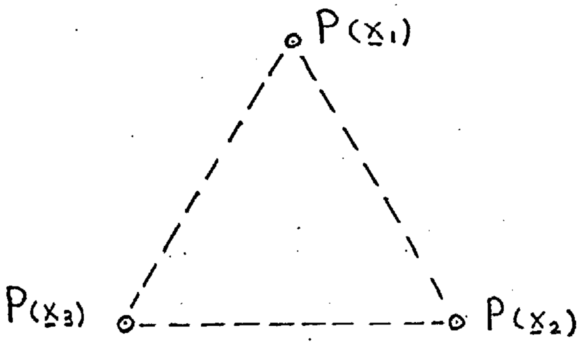


Figure 6

Three Points at Vector Coordinates $\underline{X}_1$, $\underline{X}_2$, $\underline{X}_3$, Forming an Equilateral Triangle

PARAMETRIC STUDY OF EQUATION (47)

Introduction

One method, and probably the most convenient, to attempt to measure the degree of micro-coherence of a laboratory light source would be to utilize eqn. (47) to compute the Γ_{21} to be expected at a plane illuminated by the presumed micro-coherent source, and then make comparative experimental measurements of Γ_{21} using conventional coherence measuring technique. Before attempting such an experiment it is desirable to predict the degree to which practical measurement might permit distinguishing a micro-coherent source from a classically "completely incoherent" one. To accomplish such a prediction, the behaviour of Γ_{21} as given by eqn. (47) will be studied as a function of the variables describing the source and measurement conditions. Models will be assumed for the source; these models will be chosen for their mathematical convenience, and will be physically reasonable even if not necessarily describing the conditions obtaining in a physical source.

Derivation of an Expression for Γ_{21} at a Plane Illuminated by a Gaussian Source

Assume a spherical source whose radiance is Gaussian in the radial distance from its centre, and of 1/e semi-width $\frac{1}{\sqrt{2}} s$. (The factor $\frac{1}{\sqrt{2}}$ is introduced for later mathematical convenience.) Thus the source radiance may be written as $B(|\underline{x}|) \exp\left\{\frac{-\pi 2 \underline{x}^2}{s^2}\right\}$ and its complex amplitude as $u(|\underline{x}|) = \sqrt{B(|\underline{x}|)} \exp\left\{-\pi \frac{\underline{x}^2}{s^2}\right\}$. [55]

Assume that the source micro-coherence function is also Gaussian, of 1/e semi-width σ , thus the source coherence-function is $\Gamma(\underline{x},\hat{\underline{x}}) = \exp\left\{\frac{-\pi |\underline{x}-\hat{\underline{x}}|^2}{\sigma^2}\right\}$ [56]
If there were complete correlation in phases for all radiating points (i.e., $\sigma = \infty$) , then eqn. [56] would imply that the source were completely coherent. However, for the Gaussian radiance distribution specified, there would necessarily be some degree of incoherence attributable to the lack of uniformity in intensity alone. This may be easily determined as follows:

The visibility $V = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}$, and, writing the coherence function Γ_{21} as $\Gamma_{21} = V_{21} e^{i\beta_{21}}$, the intensity is given by $I = I_1 + I_2 + 2V_{21} \sqrt{I_1 I_2} \cos \beta_{21}$. For $\cos \beta_{21} = \pm 1$ at the maximum and minimum, the visibility is

$$V = \frac{I_1 + I_2 + 2\sqrt{I_1 I_2} - I_1 - I_2 + 2\sqrt{I_1 I_2}}{I_1 + I_2 + 2\sqrt{I_1 I_2} + I_1 + I_2 - 2\sqrt{I_1 I_2}} = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2} .$$

Let $\frac{I_1}{I_2} = K$, then $V = \frac{2\sqrt{K}\, I_2}{I_2\,(1+K)} = \frac{2\sqrt{K}}{1+K}$.

As an extreme case, let $\sigma = s$, and compare the secondary reduction in coherence from the inequality in intensities to the primary coherence reduction from the lack of perfect phase correlation. For $(\underline{x} - \hat{\underline{x}}) = \frac{s}{\sqrt{\pi}}$, $K = e$, and $V = \frac{2\sqrt{e}}{e+1} = 0.89$. The primary coherence function, $\Gamma(\underline{x}-\hat{\underline{x}})$, for this separation, equals $1/e = 0.37$. Thus, even for this extreme case, the falloff in coherence due to the intensity distribution alone may be neglected.

Let the source be centered at the origin of the coordinates X, Y, Z , and let the two points 1, 2 lie at some distance symmetrically about the Z axis and on the ξ axis of subsidiary coordinates in ξ, η, ζ . Let $\xi_1 = -\xi_2$, and also let the point $\zeta = 0$ be at the origin of X, Y, Z . R_0 is the distance from the origin to P_1 or P_2 . Fig. 7 shows this arrangement. Applying eqn. (47) to this situation, for one spectral frequency v (which will be temporarily dropped from the notation), the coherence Γ_{21} is

$$\Gamma_{21} = \frac{1}{\sqrt{I_1 I_2}} \iint_{-\infty\ \underline{x}\,\hat{\underline{x}}}^{\infty} \left[e^{-\frac{\pi(\underline{x}-\hat{\underline{x}})^2}{\sigma^2}} \right] \left[\frac{\sqrt{B(|\underline{x}|)}}{R_1}\, e^{-\pi\frac{x^2}{s^2}}\, e^{-i\frac{2\pi}{\lambda}R_1} \right]$$

$$\left[\frac{\sqrt{B(|\hat{\underline{x}}|)}}{R_2}\, e^{-\pi\frac{\hat{x}^2}{s^2}}\, e^{+i\frac{2\pi}{\lambda}R_2} \right] d\underline{x}\, d\hat{\underline{x}} \, . \qquad (57)$$

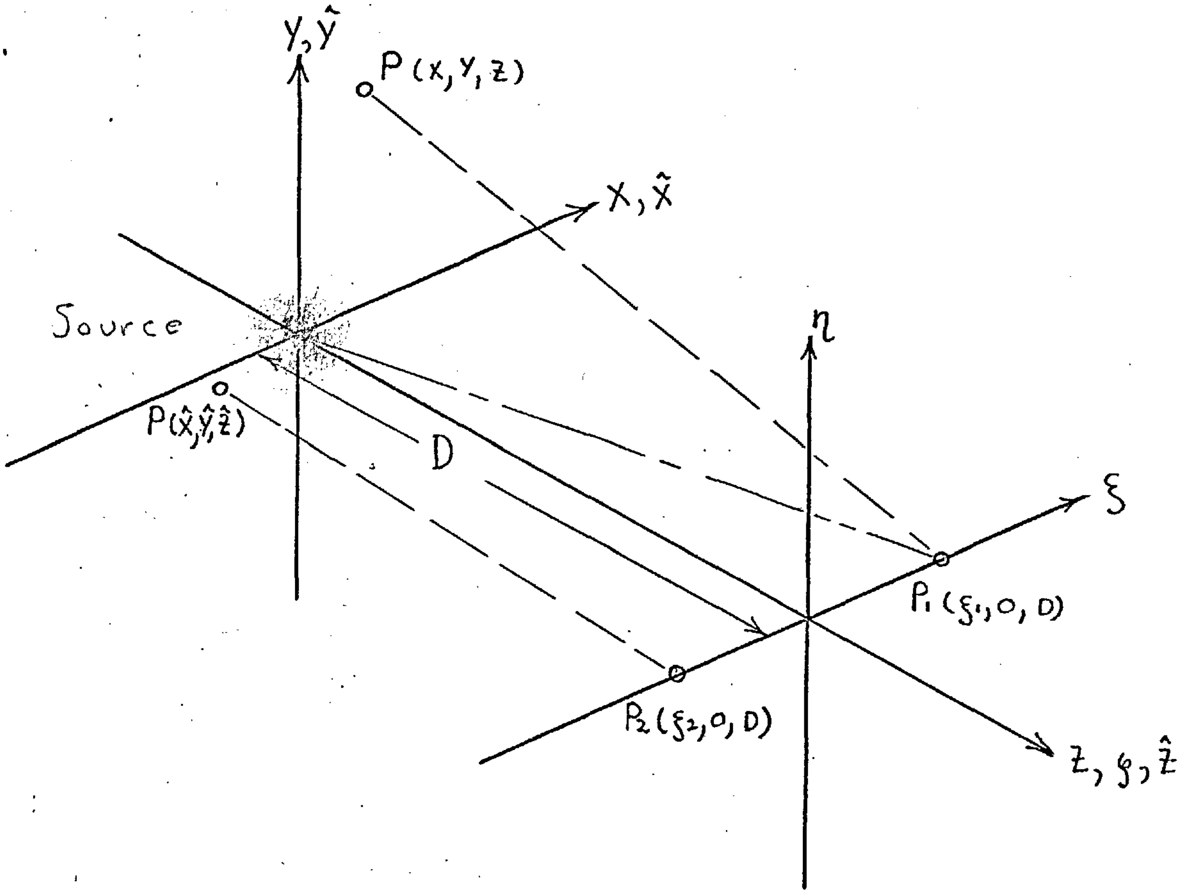


Figure 7

Two Points Illuminated by a

Gaussian Micro-Coherent Source

The infinite limits are permissible because the indefinite integral over the Gaussian is finite. The distances R_1, R_2, are as shown in Fig. 7. The variation in the factors $\frac{1}{R_1}$, $\frac{1}{R_2}$ are inconsequential, and $\frac{1}{R_1}\frac{1}{R_2}$ may be replaced $\frac{1}{D^2}$. In the exponents, R_1, R_2 may be developed as follows:

$$R_1^2 = \left[(X-\xi_1)^2 + (Y)^2 + (Z-D)^2 \right]$$ from Pythagorean geometry.

Expanding and rearranging terms gives

$$R_1^2 = X^2 + Y^2 + Z^2 - 2X\xi_1 - 2ZD + D^2 + \xi_1^2,$$

and with $R_o^2 = \xi_1^2 + D^2$,

R_1^2 may be written

$$R_1^2 = R_o^2 \left\{ 1 + \frac{X^2+Y^2+Z^2}{R_o^2} - \frac{2X\xi_1 + 2ZD}{R_o^2} \right\}.$$

For $R_o \gg (\xi - D)$ and $R_o \gg Z$, a Taylor series expansion may be used; and from the principle of stationary phase, the terms in this expansion of degree higher than two may be discarded. Thus we have

$$R_1 = R_o + \left(\frac{X^2}{2R_o} - \frac{X\xi_1}{R_o} \right) + \left(\frac{Y^2}{2R_o} \right) + \left(\frac{Z^2}{2R_o} - \frac{ZD}{R_o} \right). \quad (58)$$

Similarly,

$$R_2 = R_0 + \left(\frac{\hat{X}^2}{2R_0} - \frac{\hat{X}\xi_2}{R_0} \right) + \left(\frac{Y^2}{2R_0} \right) + \left(\frac{\hat{Z}^2}{2R_0} - \frac{\hat{Z}D}{R_0} \right) .$$

Eqn. (55) and (56) may be written in terms of the coordinate variables X, Y, Z, rather than the vector coordinate $\underline{x}$, as

$$u(X,Y,Z) = \sqrt{B(X,Y,Z)}\; e^{-\pi \frac{X^2+Y^2+Z^2}{s^2}} \tag{59}$$

and

$$\Gamma(X-\hat{X}, Y-\hat{Y}, Z-\hat{Z}) = e^{-\pi \frac{(X-\hat{X})^2 + (Y-\hat{Y})^2 + (Z-\hat{Z})^2}{\sigma^2}} . \tag{60}$$

When these expressions are used in the expression for Γ_{21} in (57), the resulting equation may be written as the product of six paired integrals, in X, $\hat{X}$; Y, $\hat{Y}$; and Z, $\hat{Z}$.

Now consider the normalization factors I_1, I_2 appearing in (57). Substitution of (55, 56) into (44) gives an expression very similar to the integral of (57); only whereas (57) is in the variables R_1, R_2, the integrals giving I_1 and I_2 will be in the variables $R_1 R_1$, and $R_2 R_2$. Thus the normalizing factor I_1 will be

$$I_1 = \iint \left[e^{-\pi \frac{(\underline{x}-\hat{\underline{x}})^2}{\sigma^2}} \right] \left[\frac{\sqrt{B(|\underline{x}|)}}{R_1}\, e^{-\pi \frac{\underline{x}^2}{s^2}}\, e^{-i \frac{2\pi}{\lambda} R_1} \right] \left[\frac{\sqrt{B(|\hat{\underline{x}}|)}}{R_1}\, e^{-\pi \frac{\hat{\underline{x}}^2}{s^2}}\, e^{+i \frac{2\pi}{\lambda} R_1} \right] d\underline{x}\, d\hat{\underline{x}} \tag{61}$$

and similarly for I_2. Noting that insofar as the Y, $\hat{Y}$, Z, $\hat{Z}$, and R_o dependence are concerned, (61) is identical to the integral of (57), and so these factors will cancel when substituted into (57), leaving only the integrals in X, $\hat{X}$ remaining. This is, of course, simply a consequence of P_1, P_2 having been chosen to lie on the ξ axis.

If (57) is written

$$\Gamma_{21} = \frac{1}{\sqrt{I_1 I_2}} Q_{21} \qquad (62)$$

then we can write the specification of Q_{21} in functional form as

$$Q_{21} = \iint_{-\infty}^{\infty} f(x-\hat{x})\, g(x)\, h(\hat{x})\, dx\, d\hat{x} \,,$$

where the three functions f, g, and h are those enclosed in the large square brackets in (57). A modified form of the triple product integral (Appendix I),

$$\iint_{x\,\hat{x}} f(x-\hat{x})\, g(x)\, h(\hat{x})\, dx\, d\hat{x} = \int_{u} F(u)\, G(-u)\, H(u)\, du \qquad (63)$$

wherein f, F; g, G; and h, H are Fourier Transform pairs, may be used to simplify (57).

The Fourier transform of the first factor is easily found by using the standard relationship

$$FT\left[e^{-\pi\frac{t^2}{a^2}}\right] = a\,e^{-\pi a^2 u^2}$$

from which the desired transform is given as

$$FT\left[e^{-\pi\frac{(x-\hat{x})^2}{\sigma^2}}\right] = \sigma\,e^{-\pi\sigma^2 u^2}. \qquad (64)$$

The Fourier transforms of the second and third factors, products of Gaussian and Fresnel type functions, may be found by contour integration in the complex plane (Appendix II); the general relationship so obtained is that

$$FT\left[e^{-\pi\alpha t^2}\,e^{-i\pi(\beta t^2+2\gamma t)}\right] = \frac{1}{\sqrt{\alpha+i\beta}}\,e^{-\pi\frac{(\gamma-u)^2}{\alpha+i\beta}}$$

$$\text{for } \alpha > 0. \qquad (65)$$

In the present case, the second two factors in eqn. (63) are given as:

$$g(x) = e^{-\pi\frac{x^2}{S^2}}\,e^{-i\frac{2\pi}{\lambda}\left(\frac{x^2}{2R_o}-\frac{x\,\xi_1}{R_o}\right)} \quad , \qquad (66)$$

$$h(\hat{x}) = e^{-\pi\frac{\hat{x}^2}{S^2}}\,e^{-i\frac{2\pi}{\lambda}\left(\frac{-\hat{x}^2}{2R_o}+\frac{\hat{x}\,\xi_2}{R_o}\right)} \qquad (67)$$

from which the following relationships are seen:

$$\alpha_1 = \frac{1}{S^2}\ , \quad \beta_1 = \frac{1}{\lambda R_o} = \beta\ , \quad \gamma_1 = \frac{-\xi_1}{\lambda R_o} = -\gamma\ ,$$

$$\alpha_2 = \frac{1}{S^2}\ , \quad \beta_2 = \frac{-1}{\lambda R_o} = -\beta\ , \quad \gamma_2 = \frac{\xi_2}{\lambda R_o} = \frac{-\xi_1}{\lambda R_o} = -\gamma\ ,$$

using the notation that $\gamma = -\gamma_1$ and $\beta = \beta_1$. Using these values in conjunction with (65), the Fourier transform of $g(x)$ and $h(\hat{x})$ are

$$G(-u) = \frac{1}{\sqrt{\alpha+i\beta}}\, e^{-\pi\frac{(-\gamma+u)^2}{\alpha+i\beta}} = \frac{1}{\sqrt{\alpha+i\beta}}\, e^{-\pi\frac{(\gamma-u)^2}{\alpha+i\beta}} \qquad (68)$$

$$H(u) = \frac{1}{\sqrt{\alpha-i\beta}}\, e^{-\pi\frac{(-\gamma-u)^2}{\alpha-i\beta}} = \frac{1}{\sqrt{\alpha-i\beta}}\, e^{-\pi\frac{(\gamma+u)^2}{\alpha-i\beta}}$$

and the term Q_{21} desired in (62) is

$$Q_{21} = \int_u \sigma\, e^{-\pi\sigma^2u^2}\, \frac{1}{\sqrt{\alpha+i\beta}}\, e^{-\pi\frac{(\gamma-u)^2}{\alpha+i\beta}}\, \frac{1}{\sqrt{\alpha-i\beta}}\, e^{-\pi\frac{(\gamma+u)^2}{\alpha-i\beta}}\, du\ .$$

Expanding and rearranging terms gives

$$Q_{21} = \frac{\sigma}{\sqrt{\alpha^2+\beta^2}} \int_u e^{-\pi\sigma^2u^2}\, e^{-\pi\frac{(\gamma^2-2\gamma u+u^2)(\alpha-i\beta)\ +(\gamma^2+2\gamma u+u^2)(\alpha+i\beta)}{\alpha^2+\beta^2}}\, du \qquad (69)$$

and collecting the terms in u^2 and u , and taking the u - independent terms outside the integral, gives

$$Q_{21} = \frac{\sigma}{\sqrt{\alpha^2+\beta^2}}\, e^{-\pi\frac{2\alpha\gamma^2}{\alpha^2+\beta^2}} \int_u e^{-\pi\left[\left(\sigma^2+\frac{2\alpha}{\alpha^2+\beta^2}\right)u^2 - 2i\left(\frac{2\beta\gamma u}{\alpha^2+\beta^2}\right)\right]}\, du. \qquad (70)$$

Using the integral relation found in Appendix III-a,

$$\int_{-\infty}^{\infty} e^{-\pi At^2}\, e^{-2\pi i\,(2Bt)}\, dt = \frac{1}{\sqrt{A}}\, e^{-\pi\frac{B^2}{A}}\ , \qquad (71)$$

with, in the present case,

$$A = \sigma^2 + \frac{2\alpha}{\alpha^2+\beta^2} \qquad \text{and} \qquad B = \frac{2\beta\gamma}{\alpha^2+\beta^2}$$

eqn. (70) gives

$$Q_{21} = \frac{\sigma}{\sqrt{\alpha^2+\beta^2}}\, e^{-\pi \frac{2\alpha}{\alpha^2+\beta^2}\gamma^2} \sqrt{\frac{\alpha^2+\beta^2}{\sigma^2(\alpha^2+\beta^2)+2\alpha}}\; e^{-\pi\left(\frac{4\beta^2\gamma^2}{(\alpha^2+\beta^2)^2}\right)\left(\frac{\alpha^2+\beta^2}{\sigma^2(\alpha^2+\beta^2)+2\alpha}\right)} . \tag{72}$$

Combining the terms both in front of and within the exp, this may be written as

$$Q_{21} = \frac{\sigma}{\sqrt{\sigma^2(\alpha^2+\beta^2)+2\alpha}}\, e^{-\pi \frac{2\alpha[\sigma^2(\alpha^2+\beta^2)+2\alpha] + 4\beta^2}{(\alpha^2+\beta^2)[\sigma^2(\alpha^2+\beta^2)+2\alpha]}\gamma^2} ,$$

and rearrangement of the terms in the exp gives

$$Q_{21} = \frac{\sigma}{\sqrt{\sigma^2(\alpha^2+\beta^2)+2\alpha}}\, e^{-\pi \frac{2\alpha\sigma^2(\alpha^2+\beta^2) + 4(\alpha^2+\beta^2)}{(\alpha^2+\beta^2)[\sigma^2(\alpha^2+\beta^2)+2\alpha]}\gamma^2} ,$$

from which cancelling the term $(\alpha^2 + \beta^2)$ in the exp gives the desired form for Q_{21},

$$Q_{21} = \frac{\sigma}{\sqrt{\sigma^2(\alpha^2+\beta^2)+2\alpha}}\, e^{-\pi \frac{2(\alpha\sigma^2+2)}{\sigma^2(\alpha^2+\beta^2)+2\alpha}\gamma^2} . \tag{73}$$

Recollecting that from (62), $\Gamma_{21} = \frac{1}{\sqrt{I_1 I_2}} Q_{21}$, I_1 and I_2 must now be evaluated. It is clear that $I_1 = I_2$, and, as with Q_{21}, consideration of only the X dependent parts of I_1 gives an equation functionally identical to (63) and a similar method of evaluation is followed. The analogues of eqn. (66), (67) will be

$$g(x) = e^{-\pi \frac{x^2}{S^2}} e^{-i\frac{2\pi}{\lambda}\left(\frac{x^2}{2R_o} - \frac{x\xi_1}{R_o}\right)} \qquad (74)$$

$$h(\hat{x}) = e^{-\pi \frac{\hat{x}^2}{S^2}} e^{-i\frac{2\pi}{\lambda}\left(\frac{\hat{x}^2}{2R_o} + \frac{\hat{x}\xi_1}{R_o}\right)} , \qquad (75)$$

giving $\alpha_1 = \frac{1}{S^2}$, $\beta_1 = \frac{1}{\lambda R_o} = \beta$, $\gamma_1 = \frac{-\xi_1}{\lambda R_o} = -\gamma$

and $\alpha_2 = \frac{1}{S^2}$, $\beta_2 = \frac{-1}{\lambda R_o} = -\beta$, $\gamma_2 = \frac{+\xi_1}{\lambda R_o} = +\gamma$,

using, as before, the notation $\gamma = -\gamma_1$ and $\beta = \beta_1$. Using these values for α, β, γ , the transforms are

$$G(u) = \frac{1}{\sqrt{\alpha+i\beta}} e^{-\pi\frac{(-\gamma+u)^2}{\alpha+i\beta}} = \frac{1}{\sqrt{\alpha+i\beta}} e^{-\pi\frac{(\gamma-u)^2}{\alpha+i\beta}} \qquad (76)$$

and

$$H(u) = \frac{1}{\sqrt{\alpha-i\beta}} e^{-\pi\frac{(\gamma-u)^2}{\alpha-i\beta}} , \qquad (77)$$

and using these values gives for the intensity I_1

$$I_1 = \int_u \sigma e^{-\pi\sigma^2 u^2} \frac{1}{\sqrt{\alpha+i\beta}} e^{-\pi\frac{(\gamma-u)^2}{\alpha+i\beta}} \frac{1}{\sqrt{\alpha-i\beta}} e^{-\pi\frac{(\gamma-u)^2}{\alpha-i\beta}} . \qquad (78)$$

Expansion and rearrangement of terms in the exp gives

$$I_1 = \frac{\sigma}{\sqrt{\alpha^2+\beta^2}} \int_u e^{-\pi\sigma^2 u^2} e^{-\pi \frac{(\gamma^2-2\gamma u+u^2)(\alpha-i\beta) + (\gamma^2-2\gamma u+u^2)(\alpha+i\beta)}{\alpha^2+\beta^2}} du,$$

and collecting the terms in u^2 and u, and taking outside the integral the parts independent of u, gives

$$I_1 = \frac{\sigma}{\sqrt{\alpha^2+\beta^2}} e^{-\pi\frac{2\alpha\gamma^2}{\alpha^2+\beta^2}} \int_u e^{-\pi\left[\left(\sigma^2 + \frac{2\alpha}{\alpha^2+\beta^2}\right)u^2 - 2\left(\frac{2\alpha\gamma u}{\alpha^2+\beta^2}\right)\right]} du . \tag{79}$$

Using the integral relation derived in Appendix III-b, that

$$\int_{-\infty}^{\infty} e^{-\pi A t^2} e^{-2\pi(2Bt)} dt = \frac{1}{\sqrt{A}} e^{+\pi\frac{B^2}{A}}, \tag{80}$$

with, in the present case, $A = \sigma^2 + \frac{2\alpha}{\alpha^2+\beta^2}$ and $B = \frac{2\alpha\gamma}{\alpha^2+\beta^2}$,

eqn. (79) gives

$$I_1 = \frac{\sigma}{\sqrt{\alpha^2+\beta^2}} e^{-\pi\frac{2\alpha\gamma^2}{\alpha^2+\beta^2}} \sqrt{\frac{\alpha^2+\beta^2}{\sigma^2(\alpha^2+\beta^2)+2\alpha}} e^{+\pi\left[\frac{4\alpha^2\gamma^2}{(\alpha^2+\beta^2)^2}\right]\left[\frac{\alpha^2+\beta^2}{\sigma^2(\alpha^2+\beta^2)+2\alpha}\right]} .$$

Combining terms in the exp gives

$$I_1 = \frac{\sigma}{\sqrt{\sigma^2(\alpha^2+\beta^2)+2\alpha}} e^{-\pi\frac{2\alpha\gamma^2[\sigma^2(\alpha^2+\beta^2)] + 4\alpha^2\gamma^2 - 4\alpha^2\gamma^2}{(\alpha^2+\beta^2)[\sigma^2(\alpha^2+\beta^2)+2\alpha]}},$$

and as before, cancellation of the term $(\alpha^2+\beta^2)$ in the exp gives

$$I_1 = \frac{\sigma}{\sqrt{\sigma^2(\alpha^2+\beta^2)+2\alpha}}\, e^{-\pi \frac{2\alpha\sigma^2}{\sigma^2(\alpha^2+\beta^2)+2\alpha}\gamma^2} . \tag{81}$$

With $\frac{1}{\sqrt{I_1 I_2}} = \frac{1}{I_1}$

the desired value of Γ_{21} from (57) is given by

$$\Gamma_{21} = \frac{Q_{21}}{I_1} = e^{-\pi \frac{4\gamma^2}{\sigma^2(\alpha^2+\beta^2)+2\alpha}} , \tag{82}$$

or, substituting for α, β, and γ, the coherence factor is

$$\Gamma_{21} = e^{-\pi\left[\frac{4}{2+(\frac{\sigma}{s})^2+(\frac{\sigma}{\lambda})^2(\frac{s}{R_0})^2}\right](\frac{s}{R_0})^2(\frac{\xi}{\lambda})^2} . \tag{83}$$

Behaviour of Eqn. (83) for Coherent and Incoherent Extremes

For a completely coherent source, $\sigma = \infty$, and eqn. (83) then becomes

$$\Gamma_{21} = e^{-\frac{1}{\infty}} = 1$$ for all γ , as expected.

For a completely incoherent source, $\sigma = 0$, and then eqn. (83) becomes

$$\Gamma_{21} = e^{-\pi \frac{2}{\alpha} \gamma^2} = e^{-\pi \frac{2 S^2 \xi^2}{\lambda^2 R_o^2}} . \qquad (84)$$

This can be compared to the result obtained using classical coherence theory (Hopkins 1967, p. 210) by a simple extension to three dimensions, using the same approach as in deriving eqn. (83) but starting from the classical equation rather than eqn. (47). Thus, from classical theory,

$$\Gamma_{21} = \frac{1}{\sqrt{I_1 I_2}} \int_{\underline{x}} \mathcal{U}_1(\underline{x}) \, \mathcal{U}_2^*(\underline{x}) \, d\underline{x} . \qquad (85)$$

The complex amplitude at point 1 from the source point $\underline{x}$ will be

$$\mathcal{U}_1(\underline{x}) = \frac{\sqrt{B(\underline{x})}}{R_1} e^{-\pi \frac{x^2}{S^2}} \quad e^{-i \frac{2\pi}{\lambda} R_1}$$

and similarly for $\mathcal{U}_2$. The situation is shown in Fig. 8 . Using these values, the product of the complex amplitudes is

$$\mathcal{U}_1(\underline{x}) \, \mathcal{U}_2^*(\underline{x}) = \frac{B(\underline{x})}{R_o^2} \; e^{-i \frac{2\pi}{\lambda} (R_1 - R_2)}$$, letting $\frac{1}{R_1 R_2} = \frac{1}{R_o^2}$

as was done before.

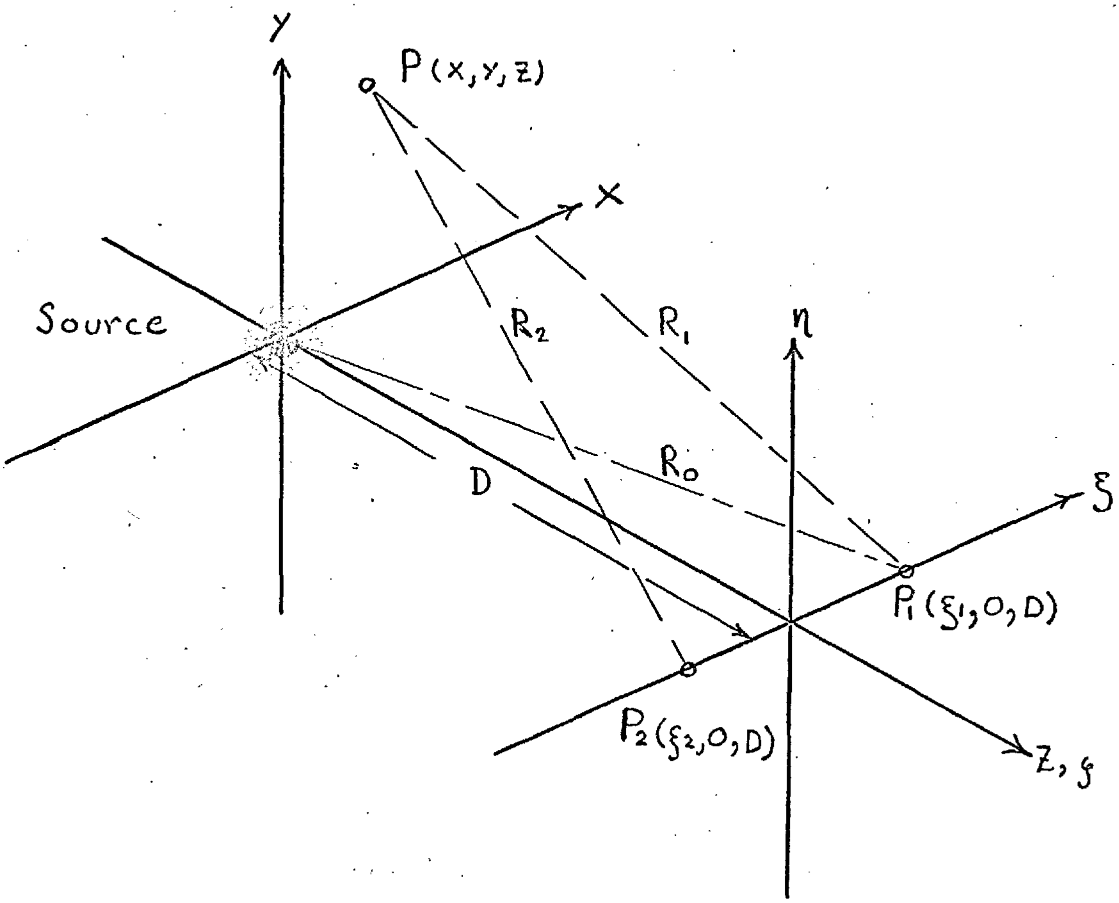


Figure 8

Two Points Illuminated by a

Gaussian Incoherent Source

The path difference $R_1 - R_2$ may be written, as before:

$$R_1^2 = (X-\xi_1)^2 + (Y)^2 + (Z-D)^2$$

or, expanding the squares,

$$R_1^2 = X^2 + Y^2 + Z^2 + \xi_1^2 + D^2 - 2X\xi_1 - 2ZD .$$

Similarly,

$$R_2^2 = X^2 + Y^2 + Z^2 + \xi_2^2 + D^2 - 2X\xi_2 - 2ZD .$$

With

$$\xi_1^2 + D^2 = R_0^2 = \xi_2^2 + D^2 ,$$

the difference $R_1^2 - R_2^2$ is

$$R_1^2 - R_2^2 = -2X\xi_1 - 2ZD + 2X\xi_2 + 2ZD$$

or, combining terms,

$$R_1^2 - R_2^2 = 2X(\xi_1 - \xi_2) .$$

Factoring $R_1^2 - R_2^2$ gives $R_1 - R_2$ as

$$R_1 - R_2 = \frac{2X(\xi_1 - \xi_2)}{R_1 + R_2} = \frac{2X\xi}{R_0}$$

higher degree terms again having been discarded. Substitution of this value into (85) gives

$$\Gamma_{21} = \frac{1}{R_o^2 \sqrt{I_1 I_2}} \int_X e^{-\pi \frac{x^2}{s^2}} e^{i \frac{\pi}{\lambda} \frac{4 x \xi}{R_o}} \, dx \int_Y e^{-\pi \frac{y^2}{s^2}} \, dy \int_Z e^{-\pi \frac{z^2}{s^2}} \, dz \, . \tag{86}$$

The standard integrals over Y and Z each equal $\frac{s}{\sqrt{2}}$, and using eqn. (71) (derived in Appendix III-a) for the integral over X , the coherence function is

$$\Gamma_{21} = \frac{1}{I_1} \frac{s^3}{R_o^2 \, 2\sqrt{2}} \, e^{-\pi \frac{2 s^2 \xi^2}{\lambda^2 R_o^2}} \, . \tag{87}$$

The normalizing factor I_1 is equal to $\frac{s^3}{R_o^2 \, 2\sqrt{2}}$, the same factor which is before the exponent of (87), just cancelling it, as shown next:

$$I_1 = \frac{1}{R_o^2} \int_X e^{-\pi \frac{x^2}{s^2}} \, dx \int_Y e^{-\pi \frac{y^2}{s^2}} \, dy \int_Z e^{-\pi \frac{z^2}{s^2}} \, dz \, ,$$

these three integrals each contributing $\frac{s}{\sqrt{2}}$ as before, thus

$$I_1 = \frac{1}{R_o^2} \left(\frac{s}{\sqrt{2}} \right)^3 = \frac{s^3}{R_o^2 \, 2\sqrt{2}} \, ,$$

leaving the coherence function

$$\Gamma_{21} = e^{-\pi \frac{2 s^2 \xi^2}{\lambda^2 R_o^2}} \, , \tag{88}$$

which agrees with eqn. (84). Thus both the classical treatment, assuming initially a completely incoherent source, and the special case of the general expression (83), yield the same result.

Numeric Evaluation of Eqn. (83)

Eqn. (83) provides a means to easily predict the effect of different degrees of micro-coherence on experimental measurements seeking to determine the extent of source micro-coherent regions; at the same time the conditions most favourable to such a measurement can be stipulated. Let such a source illuminate a distant plane, wherein lie two points P_1 , P_2 , and assume that the coherence between these two points, as a function of their separation, is to be measured by conventional experimental methods. Thus, $\Gamma_{21}(\xi)$ would be measured, and the measured values compared to the values predicted by (83) for different values of σ .

To facilitate numerical evaluation, eqn. (83) may be written as

$$\Gamma_{21}(\xi) = e^{-\pi\left(\frac{\xi}{\sigma_{21}}\right)^2} , \tag{89}$$

with $$\sigma_{21} = \sqrt{2 + \left(\frac{\sigma}{S}\right)^2 + \left(\frac{\sigma}{\lambda}\right)^2\left(\frac{S}{R_o}\right)^2}\left(\frac{\lambda R_o}{S}\right) .$$

It would be reasonable to examine the value of $\Gamma_{21}(\xi)$ at the inflexion point of the curve $\Gamma_{21}(\xi)$ *vs* ξ , as this will give the greatest change in $\Gamma_{21}(\xi)$ for a small change in σ_{21} ; the value of ξ for the inflexion point, $\xi = \frac{\sigma_{21}}{\sqrt{2\pi}}$, lies at $\frac{1}{\sqrt{2}}$ times the value of the

1/e point (Appendix IV), and thus the value of $\Gamma_{21}(\xi)$ at the inflexion point, equal to 0.606 , is sufficiently great to avoid measurement of very low visibility.

If a realistic value of R_o is chosen as $R_o = 100$ mm , and λ is chosen $\lambda = 0.5\ \mu$, it remains to choose s and ξ . Examination of eqn. (83) shows that the greatest effect on $\Gamma_{21}(\xi)$ will be noted when s is only a very few times larger than σ , and since σ may be presumed to be smaller than, say, λ , it is reasonable to take s as $s = 2\lambda$ as an initial choice. (It shall be shown that this choice is not in itself critical.) If the factors in (83) are compared, it will be seen that the factor $\left(\frac{\sigma}{\lambda}\right)^2\left(\frac{s}{R_o}\right)^2 = 10^{-10}$ and is negligible compared to the factor 2 . Neglecting this term, and setting $\sigma = 0$, gives the experimentally reasonable value for ξ of $\xi = 14.04$ mm in order to put $\Gamma_{21}(\xi)$ at the inflexion point. It is now a simple matter to evaluate the function $\Gamma_{21}(\xi)$, for $\xi = 14.04$ mm , for various values of σ , using the following data

$$R_o = 10^{-1}\text{ m}$$

$$\lambda = 0.5 \times 10^{-6}\text{ m}$$

$$S = 2\lambda = 10^{-6}\text{ m}$$

$$\xi = \frac{R\lambda}{2\sqrt{\pi}\,S} = 0.01404\text{ m}\quad\text{(for the inflexion point).}$$

The coherence function $\Gamma_{21}(\xi)$, as a function of (σ/s), is then

$$\Gamma_{21} = \exp -\pi \left[\frac{4}{\sqrt{2+(\frac{\sigma}{2\lambda})^2}} \left(\frac{1}{2\sqrt{\pi}} \right)^2 \right] = \exp -\pi \left[\sqrt{\frac{1/\pi}{2+(\frac{\sigma}{s})^2}} \right],$$

or, with $t = \sqrt{\frac{1/\pi}{2+(\frac{\sigma}{s})^2}}$,

$$\Gamma_{21} = e^{-\pi t^2}.$$

This is easily evaluated from standard tables of the Gaussian in t , and is tabulated in Table 1. A rough plot of Γ_{21} at the inflexion point, for the values which depart by a significant amount from the $\sigma = 0$ case, is shown in Fig. 9 along with the full Gaussian for $\sigma = 0$. If s were doubled or quadrupled, halving or quartering ξ would be required to stay at the inflexion point; in this case it is noted that only the ratio of $\frac{\sigma}{s}$ is important in determining the relative change in Γ_{21} . Thus, if a certain change in Γ_{21} is decided upon as the minimum measurable in an experiment, the smallest value for σ which can be detected (or measured) is determined in terms of the size of the source used. For example, if a 2 % departure of Γ_{21} from the classical value is desired, then σ must be at least 1/4 of the source size to be

	$\frac{\sigma}{S}$	$\left(\frac{\sigma}{S}\right)^2+2$	t^2	t	Γ_{21}	$\Gamma_{21}/\Gamma_{21(0)}$
$\sigma = 0$	0	2.0000	0.1590	0.399	0.606	1.000
$\lambda/8$	$1/16$	2.0039	0.1588	0.398+	0.606+	1.001
$\lambda/4$	$1/8$	2.0156	0.1580	0.398	0.607	1.002
$\lambda/2$	$1/4$	2.0625	0.1543	0.394	0.614	1.017
λ	$1/2$	2.2500	0.1414	0.376	0.641	1.058
2λ	1	3.0000	0.1061	0.326	0.716	1.182

Table 1

Computed Values for Γ_{21} at the Inflexion Point for Various σ .

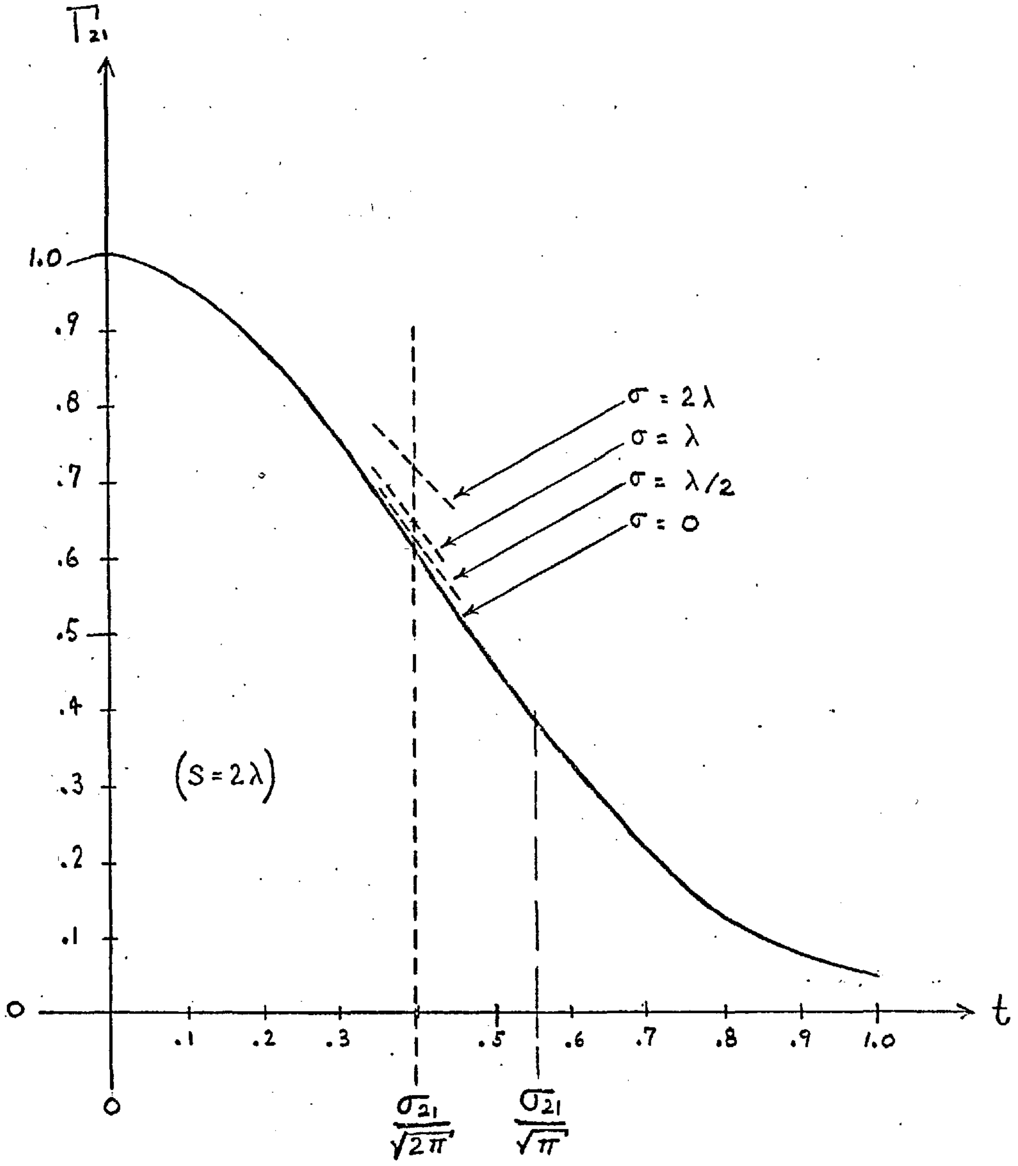


Figure 9

Plot of Γ_{21} at the Inflexion Point for Various σ.

measured properly. This clearly places stringent requirements on any experimental attempt to measure the microcoherence characteristics of a source.

Consideration of the Depth of the Source

In the preceding parts of this section, a model of a micro-coherent source was postulated to be Gaussian in intensity, the independent parameter of the Gaussian function being the distance of a given point from the centre of the source. This might be described as a "spherical Gaussian" source. It is of some interest to examine the effect of changing the depth of the source, i.e., its thickness in the Z direction, while keeping unchanged the dimensions in the transverse, or X, Y, directions. This may be conveniently done while still maintaining the Gaussian characteristics of the source, to obtain a "squashed Gaussian" or "thickened Gaussian," as the case may be. This is possible because the Gaussian function may be written in the three coordinate variables as the product of an exponential in Z and one in X, Y, thus the complex amplitude distribution function (55) may be written

$$e^{-\pi \frac{\underline{x}^2}{s^2}} = e^{-\pi \frac{z^2}{s_z^2}} \quad e^{-\pi \frac{(x^2+y^2)}{s^2}} ,$$

where s_z is the parameter determining the source depth, and s is retained to determine the source width in X, Y. It was seen in the discussion preceding eqn. (62) that, for the measurement points P_1, P_2 lying on the ξ axis normal to the Z axis, the Z - dependent parts of the integrals determining the coherence function Γ_{21} cancel to unity, and that the coherence function Γ_{21} therefore depends only on the X-dependent source characteristics. It is clear that this situation still holds regardless of whether or not the source is spherical, that is, whether or not $s_z = s$. (This will be true for any other source distribution function provided that it may be written as separable products in Z and in X, Y.)

This independence of the coherence function Γ_{21} on s_z has application if sources are considered which are unlikely to have spherical characteristics, as for example an incandescent solid tungsten filament or an excited gas discharge. The tungsten filament may be considered to radiate from only an extremely thin layer at the surface, in which case s_z would be nearly zero. The gas discharge may be considered to effectively radiate from only a thin layer at the surface of the discharge, as in cases with very strong self-absorption, leading again to a very small s_z; or the discharge may be quite transparent to its own radiation, as with a very low pressure discharge, in which case s_z might be considerably larger than s if s were limited by the construction of the apparatus.

This latter case raises an important practical question: if a small part of a large source is isolated by the apparatus to obtain the equivalent of a small source, then the foregoing statements regarding the independence of Γ_{21} on s_z do not apply. For example, if a pinhole were placed immediately before a large gas discharge tube to isolate only a small part of the discharge,as shown in Fig. 10,a quite different situation results.

It was assumed,in using in eqn. (57),the transmission functions from the source point $P(\underline{x})$ to the measurement points P_1 and P_2 as $\frac{1}{R_1} e^{-i\frac{2\pi}{\lambda}R_1}$ and $\frac{1}{R_2} e^{-i\frac{2\pi}{\lambda}R_2}$, that point $P(\underline{x})$ was in fact visible from both P_1 and P_2. As this is far from the case in this example, the transmission function would have to be drastically modified, or equivalently, the limits of integration over $\underline{x}$, $\underline{\hat{x}}$ would have to be drastically limited and be themselves dependent on P_1, P_2. The integration limits for P_1, P_2, in the integral given in (57), and for $P_1 P_1$ and $P_2 P_2$ in the normalizing integrals given by (61), would be completely different, and therefore the cancellations in Y, $\hat{Y}$, Z, $\hat{Z}$ previously enjoyed do not follow.

A more graphic way to consider the problem is illustrated in Fig. 11, in which the part of the volume of the source which can be seen by the measurement points P_1, P_2 is indicated. A small circle is drawn immediately behind the pinhole to indicate the volume Vol_{12} of the source

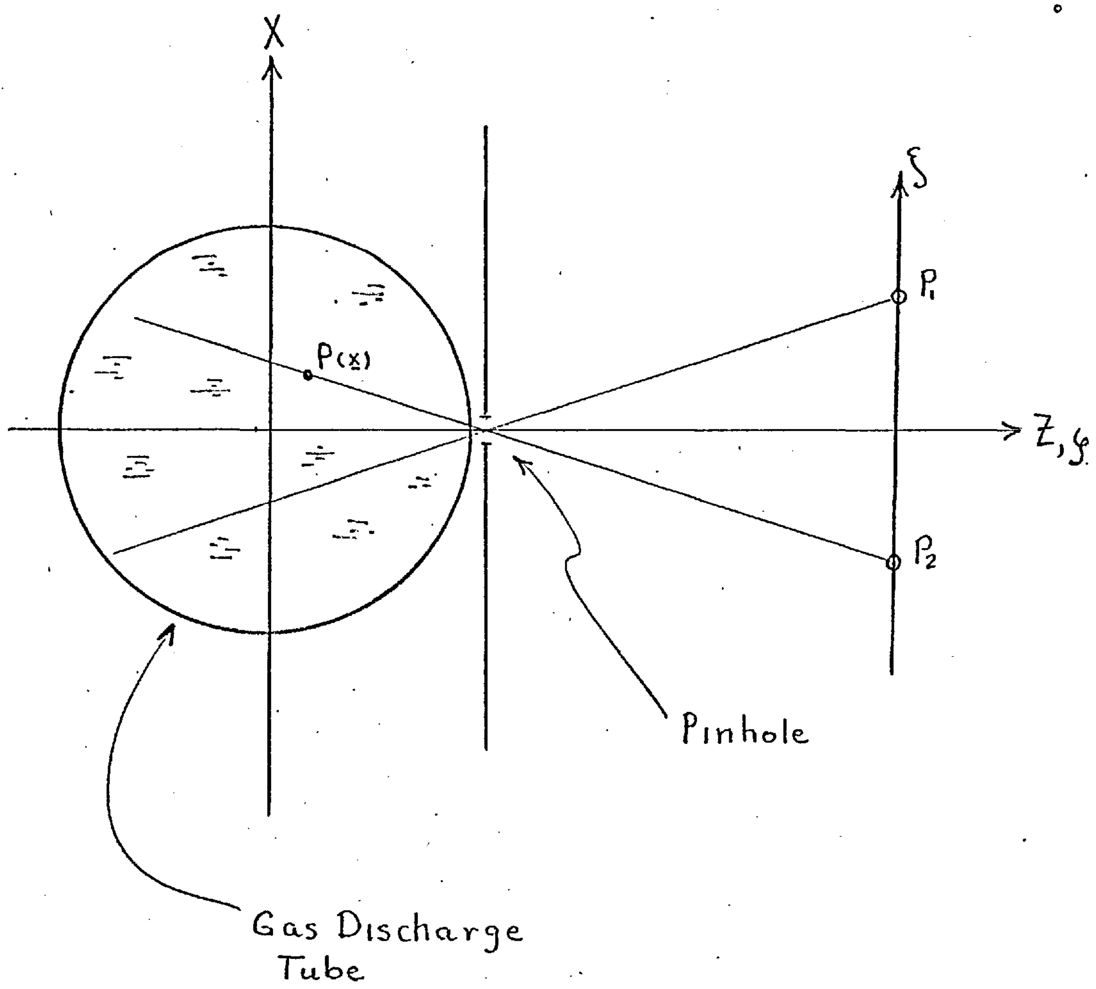


Figure 10

A Pinhole Isolating a Small Area at the Surface of a Gas Discharge Tube

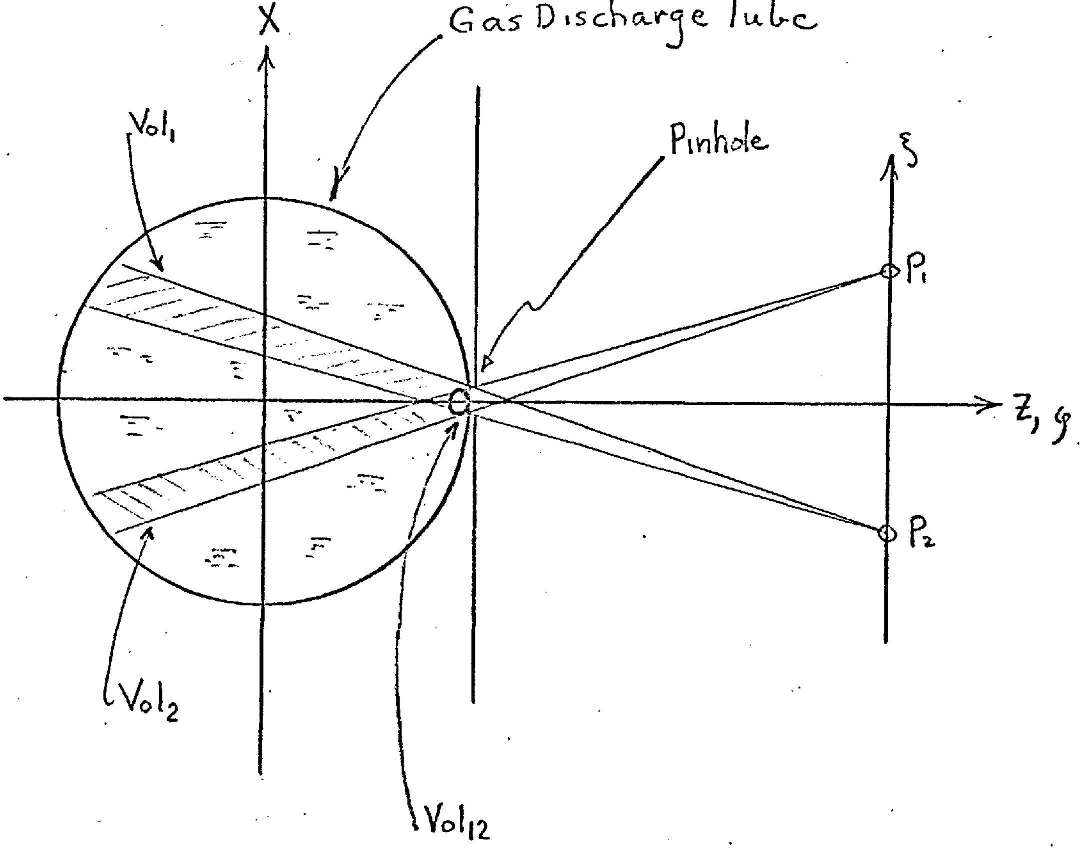


Figure 11

Volume of Overlap of the Two Volumes
Seen by Points P_1 and P_2 .

which it is desired to isolate by means of the pinhole; this volume Vol_{12} comprises the very small portions of the overlapping volumes Vol_1 and Vol_2, seen respectively from P_1 and P_2. The portion of $(Vol_1 + Vol_2)$ occupied by Vol_{12} will depend on the pinhole diameter relative to the angle subtended between P_1 and P_2. It is evident that the greater part of the light arriving at P_1 and P_2 will have arisen from the non-overlapping portions of the volumes Vol_1 and Vol_2, and that this light will very severely mask any effects arising from possible micro-coherence within the volume Vol_{12}. Thus the mathematical description in the previous paragraph is supported by this simple qualitative argument.

Consideration of Chromatic Coherence

Any experimental measurement of the micro-coherence characteristics of a classical source must necessarily utilize light of a non-zero spectral band-width; even were classical sources available which exhibited essentially perfect chromatic coherence in combination with spatial micro-coherence, or even were it possible to use essentially zero-bandwidth spectral filters or monochromating devices, it is still likely that a wider spectral bandwidth would be required in order to obtain sufficient power for reasonable experimental signal-to-noise conditions. Therefore the effects of non-zero spectral bandwidth on the foregoing formulation of spatial micro-coherence must be considered.

In developing eqn. (27) and (28), we have shown that, provided the instrumental integration time is much greater than the coherence time of the radiation, the total power at a point illuminated by a micro-coherent source may be determined by integrating with respect to spectral frequency over all of the partial monochromatic powers at the point; this is specified by eqn. (28) as

$$I(\underline{x}') = \int_{\nu} \Lambda(\nu)\, I(\nu, \underline{x}')\, d\nu \,.$$

(The factor $\Lambda(\nu)$ is the spectral sensitivity function of the detector, and is of no consequence for the present discussion.) This condition is the same as that assumed in Hopkins' (1967, p. 229) treatment of chromatic coherence, which therefore can be applied equally well to the case of a micro-coherent source as to a classical source. This treatment starts from the basic expression specifying the power at a point P illuminated by light arriving from points P_1 and P_2, as

$$I'(\bar{\nu}) = I_1'(\bar{\nu}) + I_2'(\bar{\nu}) + 2\sqrt{I_1'(\bar{\nu})\,I_2'(\bar{\nu})}\ \mathrm{Re}\left\{\Gamma_{21}(\bar{\nu})\,e^{i2\pi\bar{\nu}p(\bar{\nu})}\right\},$$

in which $\Gamma_{21}(\bar{\nu})$ is the spatial coherence between P_1 and P_2 for light of the wave-number $\bar{\nu}$, and $p(\bar{\nu})$ is the optical path difference $[P_1P] - [P_2P]$. This expression is integrated over the range $\bar{\nu} = 0$ to $\bar{\nu} = \infty$ to obtain the total power at P as

$$I' = I_1' + I_2' + 2\sqrt{I_1 I_2}\ \mathrm{Re}\{J_{21}\}$$

where J_{21} is the "total coherence function"

$$J_{21} = \frac{1}{\sqrt{I_1' I_2'}}\int_0^{\infty}\sqrt{I_1'(\bar{\nu})\,I_2'(\bar{\nu})}\ \Gamma_{21}(\bar{\nu})\,e^{i2\pi\bar{\nu}p(\bar{\nu})}\,d\bar{\nu}\,.$$

The following assumptions and definitions are now introduced:

1. $p(\bar{v})$ represents the total mean path difference from the source to the point of interference, via the two paths including P_1 and P_2.
2. The relative spectral energy distribution of the source is defined as

$$E(\bar{v}) = \sqrt{\frac{I_1'(\bar{v})\, I_2'(\bar{v})}{I_1'\, I_2'}}$$

3. The two paths via P_1 and P_2 are assumed to have the same uniform optical transmissions, and it is assumed that the path difference $p(\bar{v})$ arises solely in non-dispersive media.
4. The spatial coherence factor $V_{21}(\bar{v})$ is assumed essentially independent of wave-number; this will be true even for a very broad spectral line.
5. The wave-number origin is shifted from zero to $\bar{v}_0$, the wave-number at the centre of the spectrum.

Under these conditions, the total coherence function may be written

$$J_{21} = e^{i 2\pi \bar{v}_0 p}\, V_{21}\, K(p)$$

where $K(p)$ may be regarded as the modulus of the chromatic coherence function, and is given by the Fourier transform of the relative spectral energy distribution function as

$$K(p) = \int_{-\infty}^{\infty} E(\bar{\nu}-\bar{\nu}_o)\, e^{2\pi i(\bar{\nu}-\bar{\nu}_o)p}\, d(\bar{\nu}-\bar{\nu}_o)\ .$$

If, for example, the spatial coherence Γ_{21} at a plane illuminated by a micro-coherent source were to be measured as suggested in the foregoing sections, the total coherence factor J_{21} specifies the additional decrease in coherence to be expected from the spectral bandwidth of the source, in addition to the coherence attributable to the effects of the spatial coherence characteristics of the source. An unaccounted-for decrease in the chromatic coherence factor $K(p)$ could easily mask a slight increase in the spatial coherence visibility factor V_{21} attributable to micro-coherence effects, greatly decreasing the sensitivity of the experiment.

Consider a measurement of Γ_{21} by a classic Young's two-pinhole method, the light from the two pinholes simply falling on a screen to produce interference fringes whose visibility may be measured. As the fringe order increases from zero, the path difference factor p will increase and the chromatic coherence factor will cause a decrease in fringe visibility, as shown schematically in Fig. 12, in which $\bar{v}_o$ is the frequency at the centre of the spectrum.

If five fringes of reasonably high contrast are

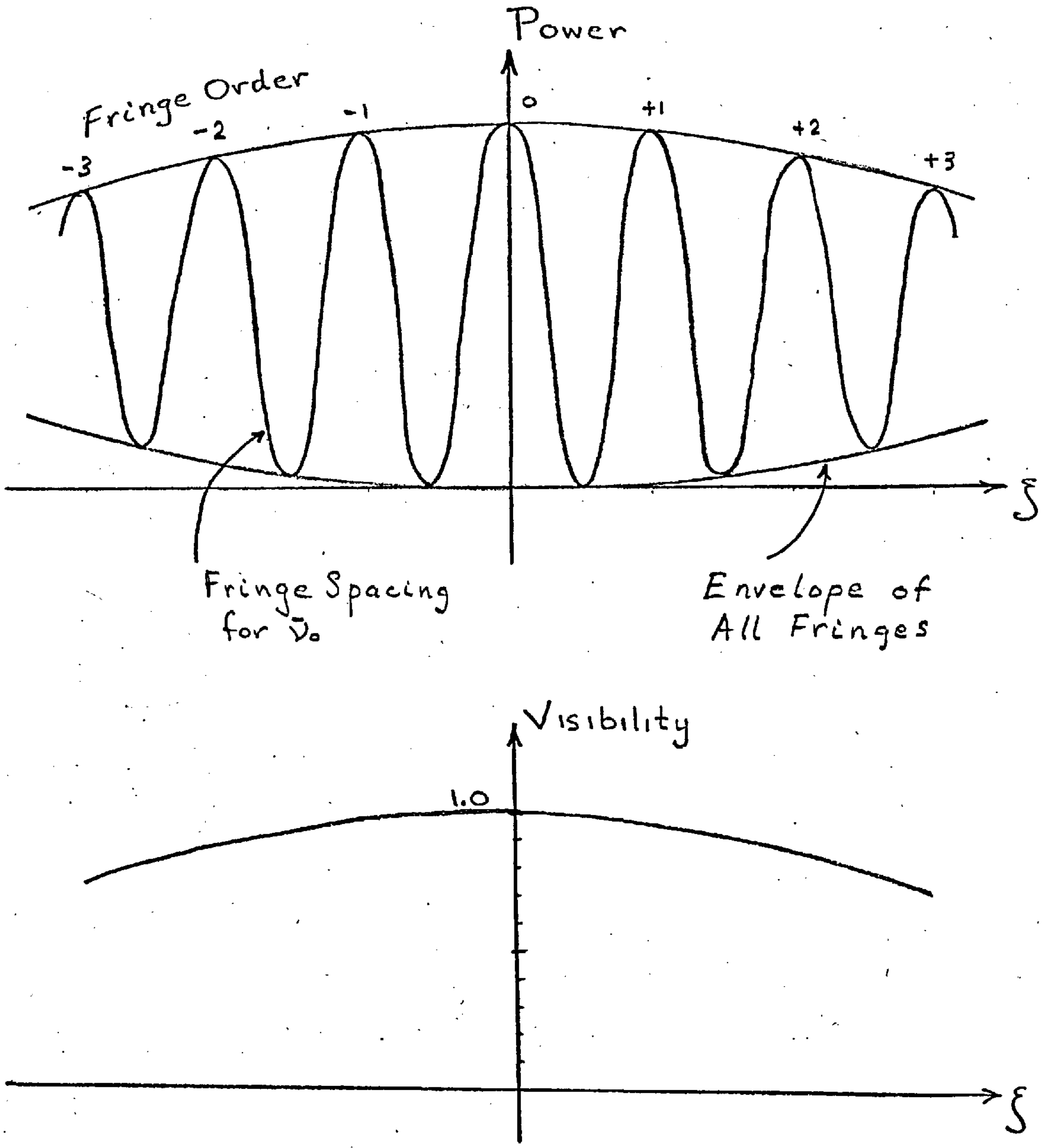


Figure 12

Power Envelope and Visibility for

Polychromatic Interference Fringes

considered to be the minimum sufficient for an accurate visibility measurement, the maximum spectral bandwidth permissible may be easily calculated. Let the chromatic coherence visibility function be required to be ≥ 0.5 at the value of p for the $\pm 2^{nd}$ fringes of $\bar{v}_0$, that is for $p = 2\lambda$, $K(p) \geq 0.5$. Again following Hopkins (1967, p. 233), assume a Gaussian profile spectral line, with a spectral energy distribution given by

$$E(\bar{\nu}-\bar{\nu}_0) = \frac{1}{\delta\bar{\nu}}\, e^{-\pi\left(\frac{\bar{\nu}-\bar{\nu}_0}{\delta\bar{\nu}}\right)^2}$$

where $\delta\bar{v}$ is the value of $\bar{v}$ for which the value of $E(\bar{\nu}-\bar{\nu}_0) = e^{-\pi/4} = 0.45$. (The factor $\frac{1}{\delta\bar{\nu}}$ is included to normalize the integrated intensity in the spectral line to unity, as required in the derivation of the total coherence factor J_{21} .) With this spectral energy distribution, the chromatic coherence function is simply $K(p) = e^{-\pi(p\delta\bar{\nu})^2}$. With $K(p) = 0.5$ and $p = 2\lambda$, we have

$$K(p) = 0.5 = e^{-\pi(2\lambda\,\delta\bar{\nu})^2},$$

from which $2(\delta\bar{v})\lambda = 0.47$, or $\lambda(\frac{1}{\lambda} - \frac{1}{\lambda_0}) = 0.24$, giving $\lambda = 0.76\,\lambda_0$. For λ_0 in the centre of the visible spectrum, use of the entire visible spectrum would be admissible. It may be remembered that an assumption was made that V_{21} was independent of $\bar{v}$, which is strictly only permissible for a spectral emission line and not for a spectral region several hundred nanometres wide; although this result is therefore somewhat optimistic as to the spectral bandwidth permissible, it does indicate that, provided the effects of chromatic coherence are properly accounted for, they do not place difficult requirements on the spectral bandwidth permitted.

SUMMARY

Without assuming any specific source emission mechanism other than the concept of the coherence length of an emitted pulse of light, an analytic specification (26) has been derived which describes the coherence between two elements of an ordinary thermally radiating source. This coherence factor is specified in terms of the time-averaged complex-amplitude cross-product of the spectra from the two radiating elements, and the legitimacy of representing the radiation by these infinite monochromatic spectral waves and by a scalar theory is substantiated. It is shown that for the usual case involving low or moderate resolution optics, the standard classical coherence theory is entirely correct and applicable.

It has been shown how this coherence factor may be used as a weighting function in integrating (28) over the complex amplitudes arriving at an observation point to determine the total power arriving there. This expression reduces to the standard forms for a perfectly coherent and perfectly incoherent source, provided that the radiation density from the source is properly normalized. If a two-point source is assumed, (28) reduces to the standard Young's fringe expression, as expected. It has been shown that the source micro-coherence function must be real.

Equation (28) has been used to derive an expression (47) giving the degree of coherence between two points at an intermediate plane illuminated by a micro-coherent source, in terms of the source micro-coherence and the complex optical paths between the source and the intermediate plane. This expression has been applied to a spherical source postulated to have a Gaussian radiance distribution and a micro-coherence function which is also Gaussian; this treatment indicates the influence of experimental conditions on the feasibility of measurement, and points to potential experimental problems. A parametric study of this situation has shown that in order to obtain reasonable experimental measurements of source micro-coherence, the source must itself be extremely small, probably only a few wavelengths in extent. It is also shown, however, that the requirements on chromatic coherence are not severe, a factor which could help alleviate the low power to be expected from such a small source.

APPENDIX I - The "Triple Product Integral"

$$\iint_{-\infty}^{\infty} f(-x-\hat{x})\, g(x)\, h(\hat{x})\, dx\, d\hat{x} =$$

$$= \int_{-\infty}^{\infty} F(u)\, G(u)\, H(u)\, du \,.$$

where f,F; g,G; and h, H are Fourier transform pairs. The integral to be developed,

$$\iint_{x\,\hat{x}} f(-x-\hat{x})\, g(x)\, h(\hat{x})\, dx\, d\hat{x} \;,$$

may be rewritten with the function g(x) written as the Fourier transform of its Fourier transform G(u) , as

$$\iint_{x\,\hat{x}} f(-x-\hat{x})\, h(\hat{x}) \left[\int_{u} G(u)\, e^{-i2\pi x u} du \right] dx\, d\hat{x} \,.$$

Upon rearrangement of the terms and order of integration, we have

$$\int_{u} G(u) \iint_{x\,\hat{x}} \left\{ e^{-i2\pi u(\hat{x}+x)} f[-(x+\hat{x})]\, dx \right\} \left\{ h(\hat{x})\, e^{+i2\pi u \hat{x}} d\hat{x} \right\} du$$

and each of the terms in curley brackets { } is seen to be one of the desired Fourier transforms, giving the desired final form as

$$\int_u G(u)\,F(u)\,H(u)\,du = \int_u F(u)\,G(u)\,H(u)\,du.$$

It is more often desired to have the triple integral in the form:

$$\iint_{x\,\hat{x}} f(x-\hat{x})\,g(x)\,h(\hat{x})\,dx\,d\hat{x} = \int_u F(u)\,G(-u)\,H(u)\,du\ .$$

To evaluate this modified form of the triple integral, start with the standard form

$$\iint_{x\,\hat{x}} f(-x-\hat{x})\;g(x)\;h(\hat{x})\;dx\;d\hat{x} = \int_u F(u)\;G(u)\;H(u)\;du$$

and make a change of variable, letting $t = -x$; and $dx = -dt$. In the new variables, the standard triple product integral relationship becomes

$$\int_t\int_{\hat{x}} f(t-\hat{x})\,g(-t)\;h(\hat{x})\,dt\,d\hat{x} = \int_u F(u)\,G(u)\;H(u)\;du.$$

Using the standard relationship that $FT[g(-t)] = G(-u)$, the triple product integral is now

$$\iint_{t\,\hat{x}} f(t-\hat{x})\,g(+t)\,h(\hat{x})\,dt\,d\hat{x} = \int_u F(u)\,G(-u)\,H(u)\,du.$$

With a second variable change, with t = x , we now have

$$\int_{x}\int_{\hat{x}} f(x-\hat{x})\, g(x)\, h(\hat{x})\, dx\, d\hat{x} = \int_{u} F(u)\, G(-u)\, H(u)\, du$$

as desired.

APPENDIX II

To show that for $\alpha > 0$,

$$\int_{-\infty}^{\infty} e^{-\pi\alpha t^2} e^{-i\pi(\beta t^2+2\gamma t)} e^{i2\pi u t}\, dt = \frac{1}{\sqrt{\alpha+i\beta}}\, e^{-\pi\frac{(\gamma-u)^2}{\alpha+i\beta}} ,$$

let the integral to be evaluated be denoted A , thus

$$A = \int_{-\infty}^{\infty} e^{-\pi[(\alpha+i\beta)t^2 + (2i\gamma-2iu)t]}\, dt ,$$

which, after completing the square, becomes

$$A = \int_{-\infty}^{\infty} e^{-\pi\left\{(\alpha+i\beta)\left[t + \frac{i(\gamma-u)}{\alpha+i\beta}\right]^2 - (\alpha+i\beta)\left[\frac{i(\gamma-u)}{\alpha+i\beta}\right]^2\right\}}\, dt .$$

Taking the term not depending on t outside the integral, and taking $(\alpha + i\beta)$ inside the square, gives

$$A = e^{-\pi\frac{(\gamma-u)^2}{\alpha+i\beta}} \int_{-\infty}^{\infty} e^{-\pi\left[\sqrt{\alpha+i\beta}\, t + \frac{i(\gamma-u)}{\sqrt{\alpha+i\beta}}\right]^2}\, dt$$

or, letting the remaining integral be denoted by B ,

$$A = e^{-\pi\frac{(\gamma-u)^2}{\alpha+i\beta}}\, B .$$

To evaluate the integral B , let

$$z = \sqrt{\alpha + i\beta}\, t + \frac{i(\gamma - u)}{\sqrt{\alpha + i\beta}}$$

in the complex plane, with $dt = \frac{1}{\sqrt{\alpha+i\beta}}\, dz$. Then

$$B = \frac{1}{\sqrt{\alpha+i\beta}} \int e^{-\pi z^2}\, dz ,$$

the integral to be taken along the line

$$z = \sqrt{\alpha + i\beta}\, t + \frac{i(\gamma - u)}{\sqrt{\alpha + i\beta}} .$$

A contour Γ may be taken to include part of this line, the Real axis, and two connecting circular arcs of radius R about the point t_o , the intercept of the line on the Real axis (Fig. A-II-1). The contour sections are shown as A, B, C, and D . Let θ be the angle between a point on this contour and the real axis, about t_o .

The integral along this contour Γ includes in turn each of the four parts and is equal to zero, thus

$$\int_{\Gamma} = \int E + \int C_+ - \int D + \int C_- = 0$$

or, solving for the integral along z ,

$$\int E = \int_{t_o - R}^{t_o + R} e^{-\pi t^2}\, dt - \int C_+ - \int C_- .$$

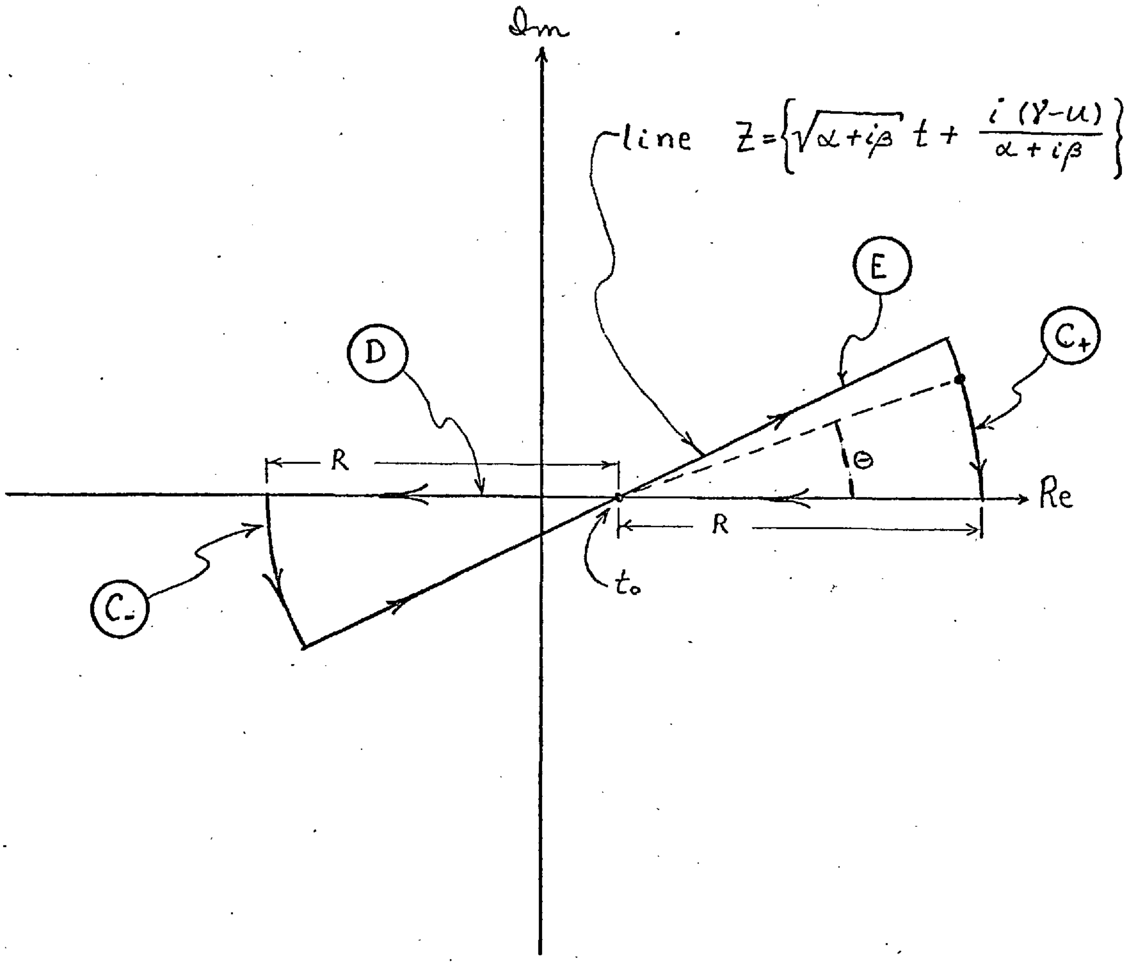

Im
line $Z = \left\{ \sqrt{\alpha + i\beta}\, t + \frac{i(\gamma - u)}{\alpha + i\beta} \right\}$
E
C_+
D
R
Θ
Re
R
C_-
t_o

Figure A-II-1

The Contour around Γ in the Complex Plane

The integrals along the circular arcs, C_+ and C_-, may be written in circular coordinates with $z = t_0 \pm R e^{i\theta}$ and $dz = \pm R e^{i\theta} i \, d\theta$, as:

$$\int_{C_+} e^{-\pi t^2} dt = \int_0^{\theta_0} e^{-\pi(t_0^2 + 2t_0 R e^{i\theta} + R^2 e^{i2\theta})} \; R e^{i\theta} \, d\theta$$

which becomes, upon expansion and rearrangement of terms,

$$\int_0^{\theta_0} e^{-\pi(t_0^2 + 2t_0 R\cos\theta + R^2\cos 2\theta)} \; e^{-i\pi(2t_0 R\sin\theta + R^2\sin 2\theta)} \; R e^{i\theta} i \, d\theta.$$

In this integrand, the terms

$$e^{-i\pi(2t_0 R\sin\theta + R^2\sin 2\theta)} \quad e^{i\theta}$$

lie inside or on the unit circle in the complex plane, and therefore have a modulus = 1 . If now the radius R is let to approach infinity, the term remaining is dominated by the factor $R^2\cos^2 2\theta$, and approaches zero, as long as $\cos 2\theta > 0$, requiring that $0 < |2\theta| < \pi/2$ or $0 < |\theta| < \pi/4$. The value of θ does in fact lie between 0 and $\pi/4$ provided $\alpha > 0$, as shown below.

For $\alpha > 0$, the complex number $(\alpha + i\beta)$ must lie to the right of, and not on, the imaginary axis; the imaginary number $\pm\sqrt{\alpha+i\beta}$ will therefore lie inside, and not on, the lines through the origin at an angle of $\pm\,\pi/4$. This is more easily seen in circular coordinates (Fig. A-II-2): $z = Re^{i\phi}$, and for $\alpha > 0$, $-\frac{\pi}{2} < \phi < \frac{\pi}{2}$; so $\sqrt{z} = Re^{i\frac{\phi}{2}}$, with $-\frac{\pi}{4} < \phi < \frac{\pi}{4}$. (It may also be noted that the integral $\int e^{-\pi z^2} dz$ is analytic for all z except $z = \infty$; it is therefore necessary to consider limits as $z \to \infty$ rather than simply values for $z = \infty$.) As $R \to \infty$, the integral along D approaches $\int_{-\infty}^{\infty} e^{-\pi t^2} dt$, which equals 1 . Thus, as $R \to \infty$, the integral $\int E$ approaches the integral $\int_{-\infty}^{\infty} e^{-\pi z^2} dz$, and in the limit as $R \to \infty$, we have

$$\int E = \int D - \int C_+ + \int C_- = 1 - 0 - 0 = 1 .$$

Thus $B = \frac{1}{\sqrt{\alpha+i\beta}}$, and the desired integral A is seen to be

$$A = \frac{1}{\sqrt{\alpha+i\beta}}\, e^{-\pi\frac{(\gamma-u)^2}{\alpha+i\beta}} .$$

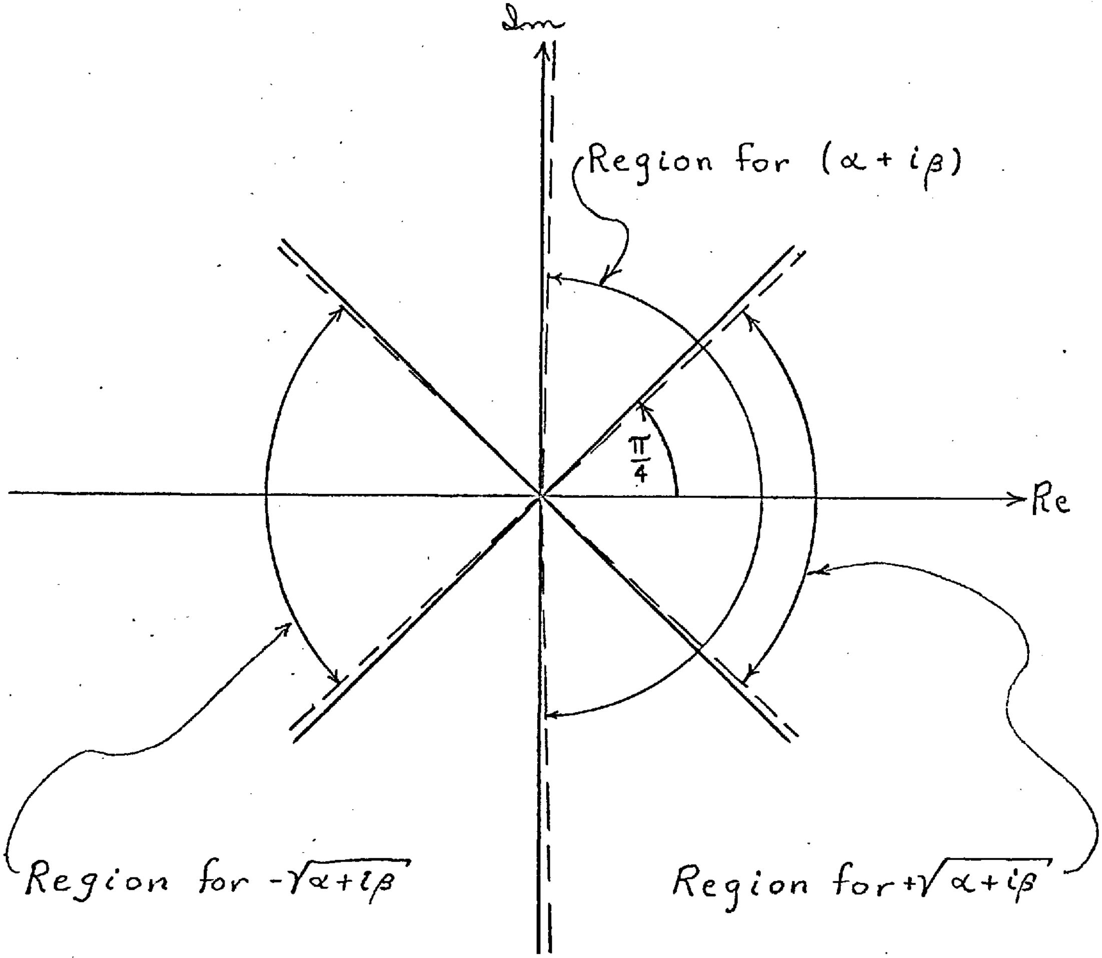

Im
Region for $(\alpha + i\beta)$
$\frac{\pi}{4}$
Re
Region for $-\sqrt{\alpha+i\beta}$
Region for $+\sqrt{\alpha+i\beta}$

Figure A-II-2

For $\alpha > 0$, the Line $\sqrt{\alpha+i\beta}$ lies within the Octant ± $\pi/4$ about the Real Axis

APPENDIX III - Evaluation of Two Integrals

(a) To show that $\int_{-\infty}^{\infty} e^{-\pi A t^2} e^{-i\pi 2Bt}\, dt = \frac{1}{\sqrt{A}}\, e^{-\pi \frac{B^2}{A}}$.

First, complete the square and take outside the integral the parts not depending on t :

$$\int_{-\infty}^{\infty} e^{-\pi(At^2+i2Bt)}dt = \int_{-\infty}^{\infty} e^{-\pi[(\sqrt{A}t+\frac{iB}{\sqrt{A}})^2+\frac{B^2}{A}]}\, dt =$$

$$= e^{-\pi\frac{B^2}{A}} \int_{-\infty}^{\infty} e^{-\pi(\sqrt{A}t+\frac{iB}{\sqrt{A}})^2}\, dt .$$

The remaining integral must be examined in the complex plane, where, with $z = \sqrt{A}t + \frac{iB}{\sqrt{A}}$ and $dz = \sqrt{A}\, dt$, the integral may be written

$$\frac{1}{\sqrt{A}} \int_{z=-\infty+\frac{iB}{\sqrt{A}}}^{z=+\infty+\frac{iB}{\sqrt{A}}} e^{-\pi z^2}\, dz \quad ;$$

the integral is along the line $z = \sqrt{A}\, t + \frac{iB}{\sqrt{A}}$, which is parallel to the Real axis as shown in Fig. A-III. The point (X,0) , denoted simply as X , is a running point on the Real axis. A contour Γ may be taken through points A, B, C, D as shown on the Figure, the integral around the contour Γ being the sum of the four parts of

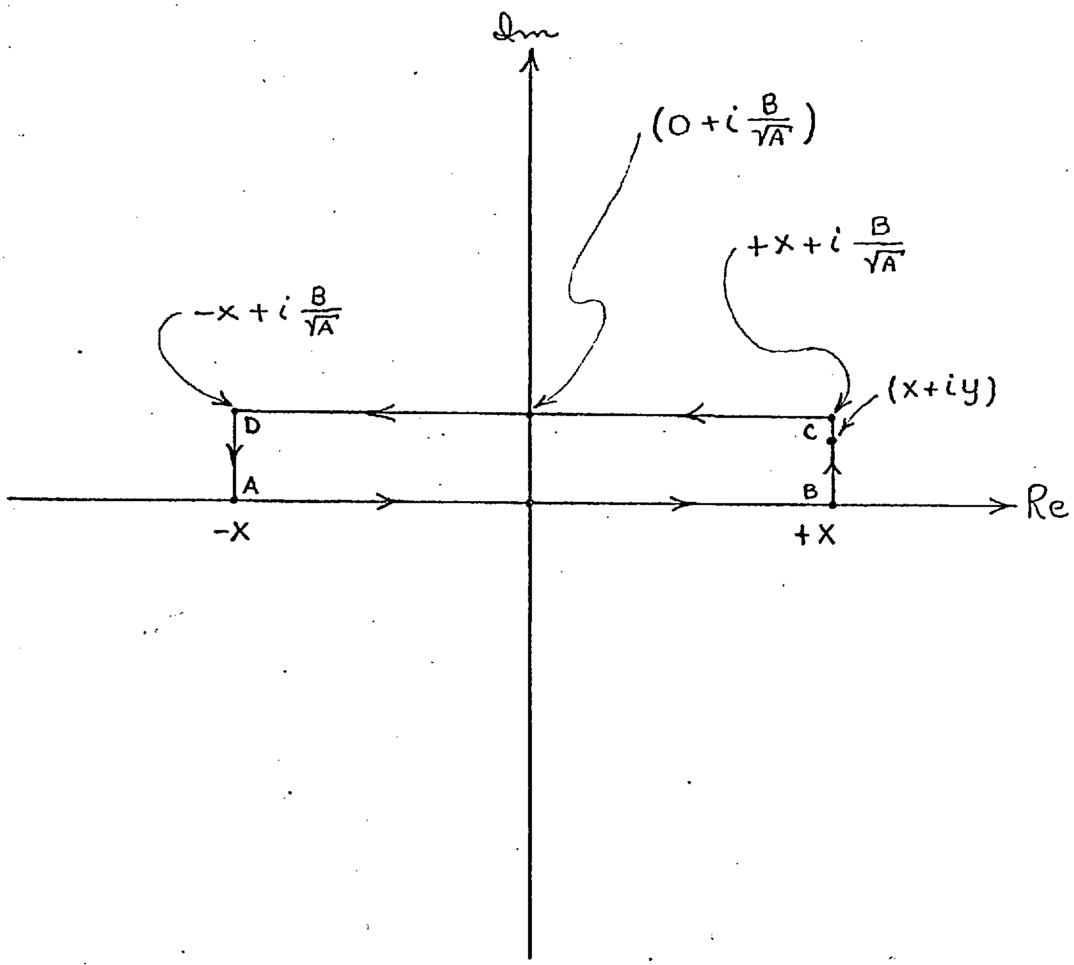


Figure A-III

An Integration Contour in the Complex Plane the contour, thus $\int_{T} = \int_{A}^{B} + \int_{B}^{C} + \int_{C}^{D} + \int_{D}^{A}$. There are no singularities within or on this contour, so the contour integral is equal to zero, and we may write

$$\int_{D}^{C} = \int_{A}^{B} + \int_{B}^{C} - \int_{A}^{D} .$$

Consider the integral $\int_{B}^{C}$, along a path between B and C which may be written as $Z = X + iy$ (with $dz = idy$) , giving the integral as

$$\int_{B}^{C} = \int_{y=0}^{y=\frac{B}{\sqrt{A}}} e^{-\pi(X+iy)^2}\, i\,dy ,$$

or expanding and re-arranging terms,

$$\int_{B}^{C} = \int_{y=0}^{y=\frac{B}{\sqrt{A}}} e^{-\pi(X^2+2Xiy-y^2)}\, i\,dy = \int_{y=0}^{y=\frac{B}{\sqrt{A}}} e^{-\pi(X^2-y^2)}\, e^{-\pi i 2Xy}\, i\,dy.$$

The modulus of the factor $e^{-\pi i 2Xy}$ is equal to unity, and if X is now let to approach infinity, the X^2 factor swamps the y^2 factor in the term $e^{-\pi(X^2-y^2)}$, and the value of the integral approaches zero. It is clear that the integral $\int_{D}^{A}$ will similarly approach zero. This leaves that

$$\int_{D}^{C} = \int_{A}^{B} \quad as \quad X \to \infty \; ;$$

i.e., that the value of the integral of the complex variable

$$\int_{-\infty+iy}^{+\infty+iy} e^{-\pi z^2} dz$$

equals the value of the integral over the Real part of Z ,

$$\int_{-\infty}^{\infty} e^{-\pi t^2} dt \; .$$

This latter integral has the standard solution

$$\int_{-\infty}^{\infty} e^{-\pi t^2} dt = 1 \quad ,$$

or

$$\int_{-\infty}^{\infty} e^{-\pi K t^2} dt = \frac{1}{\sqrt{K}} \quad .$$

Applying this to the present integral, where $K = A$, the desired result is obtained that

$$\int_{-\infty}^{\infty} e^{-\pi A t^2} e^{-i\pi 2 B t} dt = \frac{1}{\sqrt{A}} e^{-\pi \frac{B^2}{A}} \; .$$

(b) <u>To show</u> that $\displaystyle\int_{-\infty}^{\infty} e^{-\pi A t^2} e^{-\pi 2 B t} dt = \frac{1}{\sqrt{A}} e^{+\pi \frac{B^2}{A}}$.

As before, completing the square and taking outside the integral the parts not depending on t :

$$\int_{-\infty}^{\infty} e^{-\pi A t^2} e^{-\pi 2 B t} dt = \int_{-\infty}^{\infty} e^{-\pi [(\sqrt{A} t + \frac{B}{\sqrt{A}})^2 - \frac{B^2}{A}]} dt =$$

$$= e^{+\pi \frac{B^2}{A}} \int_{-\infty}^{\infty} e^{-\pi (\sqrt{A} t + \frac{B}{\sqrt{A}})^2} dt \; .$$

Again using the standard result that $\int_{-\infty}^{\infty} e^{-\pi K t^2} dt = \frac{1}{\sqrt{K}}$, the desired result is

$$\int_{-\infty}^{\infty} e^{-\pi A t^2} e^{-\pi 2 B t} \, dt = \frac{1}{\sqrt{A}} e^{+\pi \frac{B^2}{A}} .$$

APPENDIX IV - Inflexion Point of the Gaussian

The Gaussian function $e^{-\pi\frac{t^2}{\sigma^2}}$ will have its steepest slope at the value of t for which the first derivative is zero - i.e., the zero of the second derivative, or inflexion point. This is easily found; let f(t) denote the Gaussian, $f(t) = e^{-\pi\frac{t^2}{\sigma^2}}$.

The first derivative, giving the slope, is

$$f'(t) = \frac{-\pi 2t}{\sigma^2} e^{-\pi\frac{t^2}{\sigma^2}},$$

and the second derivative, giving the curvature, is

$$f''(t) = \left[\left(\frac{-\pi 2t}{\sigma^2}\right)^2 - \frac{2\pi}{\sigma^2}\right] e^{-\pi\frac{t^2}{\sigma^2}}.$$

Setting the second derivative equal to zero,

$$f''(t) = 0 = \frac{4\pi^2 t^2}{\sigma^4} - \frac{2\pi}{\sigma^2}$$

gives the value of t for the inflexion point as

$t^2 = \frac{\sigma^2}{2\pi}$ and $t = \frac{\sigma}{\sqrt{2\pi}}$ at the inflexion point.

This may be compared to the value of t at the point where f(t) = 1/e : thus for $e^{-\pi\frac{t^2}{\sigma^2}} = e^{-1}$,

we have $t = \frac{\sigma}{\sqrt{\pi}}$ at the 1/e point.

BIBLIOGRAPHY

Beran, M., and Parrent, G. (1964), Theory of Partial Coherence. Englewood

Born, M., and E. Wolf (1959), Principles of Optics. Pergamon Press Ltd.

Bridgman, P. (1927), The Logic of Modern Physics, Macmillan

Davenport, W., and Root, W. (1958), An Introduction to the Theory of Random Signals and Noise. McGraw-Hill

Dirac, P. (1958 Fourth Edition), The Principles of Quantum Mechanics. Oxford, Clarendon Press

Heavens, O.S. (1954), Optical Masers. Methuen and Co. Ltd.

Hopkins, H.H. (1967), The Theory of Coherence and Its Applications, in Advanced Optical Techniques, editor: A.C.S. van Heel. Amsterdam: North Holland Publishing Co.

Hopkins, H.H. (1951), The concept of partial coherence in optics, Proc. Roy. Soc., 208A, p. 263

Mandel, L., and G. Magyar (1963), Nature, 198, p. 255

Mandel, L. (1964), Quantum Theory of Interference Effects Produced by Independent Light Beams, Phys. Rev. 134, p. A10-15

Stone, J. (1963), Radiation and Optics. McGraw-Hill

ACKNOWLEDGEMENTS

I wish to thank three friends who have contributed more than they guess to the successful completion of this thesis: Larry Mertz, who initiated me to "Fourier Optics"; Adolf Lohmann, who broadened and deepened my interest in this field; and Sumner Davis, whose unselfish advice over the past dozen years has been invaluable.

I wish to thank my employer, the International Business Machines Corporation, for their financial support during a part of my postgraduate work.

I wish to thank H. H. Hopkins for suggesting this problem, for guidance in developing it, for suggestions substantially improving this manuscript, and for a most stimulating and rewarding association.

JAPANESE JOURNAL OF APPLIED PHYSICS VOL. 4, SUPPLEMENT I, 1965
PROCEEDINGS OF THE CONFERENCE ON PHOTOGRAPHIC AND SPECTROSCOPIC OPTICS, 1964

Superresolution Image-Forming Systems for Objects with Restricted Lambda Dependance

J. D. ARMITAGE, A. LOHMANN, and D. P. PARIS
IBM Corporation, San Jose, California, U.S.A.

When talking about "resolution" we refer to the fact that optical instruments usually act as low pass filters, letting pass only frequencies $|R| \leq R_A$. The cutoff frequency R_A, when induced by diffraction on the lens aperture, depends on wavelength λ and stop number æ such that $R_A = 1/\lambda æ$. This is true for any incoherently illuminated object. However, if only certain special classes of objects are admitted, one can improve the resolution considerably. Then the image-forming apparatus needs some parts in additon to the lens, as for example spectral prisms, moving masks or Wollaston prisms. Such apparatus, which we will call "superresolution" devices, have been built by several investigators. The following special object classes were used: restricted λ dependence[1] restricted time dependence,[2] nonbirefringence,[3] relief structure.[4]

In two cases one can go so far that "resolution" of the intrinsic optical system (the lens) is reduced even to zero. That is, the lens can be replaced for example by a ground glass or by a scrambling light pipe, which can transmit only the temporal and spectral contents of a beam of light, but no spatial structure. Then the spatial structure of the object has to be encoded either as spectral structure[1] of the light beam or as temporal variations.[5]

This paper deals with spectral encoding and transmission through a light pipe. After explaning the basic idea, which has been independently conceived by Downes (1918), Lindenblad (1948) and Kartashev (1960), we will show how one can overcome two serious drawbacks: restriction to *one*-dimensional objects; very low light throughput. Then we will describe several modifications, some of them admitting a larger variety of objects, others performing certain analog computations.

I. Basic Principle

The first prism spectroscope displays a continuous spectrum in the object plane. The object, here two black lines, blocks out two particular wavelengths. The light is collected and transmitted without any spatial resolution to the second spectroscope, which "decodes" the signal by displaying the transmitted spectrum. The two missing wavelengths give rise to two dark stripes, which constitute the "image." The color is irrelevant. Obviously this method works only for *one*-dimensional x-dependent amplitude objects, being black-grey-white. The light throughput is very low due to the entrance slits of the two spectroscopes.

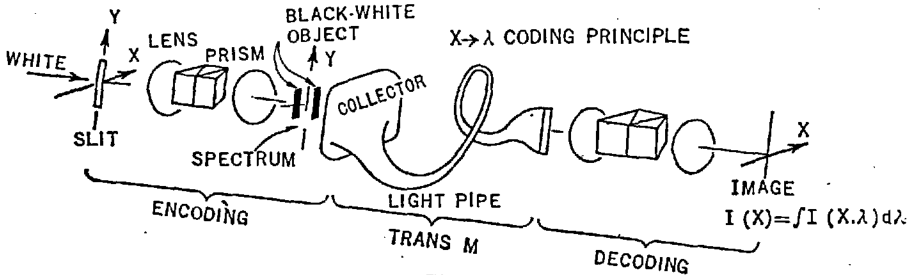


Fig. 1.

II. Extension to Two Dimensions

While the x-dependence of the object is λ encoded, the y-dependence can be time encoded,[2),5)] for example by moving two masks in both object and image plane in y-direction the way it was described in the preceding paper by Lukosz.

III. Increased Light Throughput

Replace the entrance slits $\delta(x)$ by two identical masks $M(x)$. The optical transfer function of the total system is essentially the spatial power spectrum $|\tilde{M}(R)|^2$ of the masks. Fresnel zone plates and pseudo noise pattern, both known from spectrometry and Radar as "sharp autocorrelators," are suitable masks.

IV. Extension to Colored Objects

With masks according to III also colored objects are allowed, if their transmittance $I_0(X, \lambda)$ changes slowly as a function of λ.

V. Extension to Phase Objects

Many ways are known to produce an amplitude image from a pure phase object: interference microscopy, phase contrast, Schlieren methods. These methods can be somehow incorporated into the spectral encoding part in order to extend the applicability of encoding to phase objects.

VI. Interference Filter instead of Prism

While a prism displays a spectrum in the form $\lambda \rightarrow X$ an interference filter in a divergent beam performs as $\lambda \rightarrow r$. Since most objects will depend on *both* polar coordinates r and φ an additional encoding $\phi \rightarrow t$ is needed. Two sector discs $M(\varphi)$ synchroneously spinning in object and image plane will accomplish this.

VII. Correlation and Convolution Experiments

So far we described several ways of image transmission, now we will show how the basic principle of λ encoding can be applied to something quite different: optical analog computation; in particular correlation and convolution. The first spectroscope in Fig. 2 is used to create shift $a\lambda$ of the function g, replacing now the entrance slit.

$$\iint f\,(X',Y)g\,(X'+a\lambda,Y')\,dx'dy' = \phi\,(a\lambda,o)$$

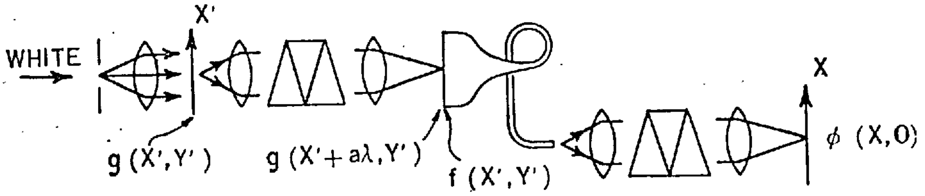


Fig. 2. Optical generation of correlation $\phi(x, y)$.

The multiplication takes place where the second function $f(x', y')$ is located as a mask. The integration is performed by the collector. The conversion of the correlation parameter (shift) $a\lambda$ from a wavelength into a geometrical coordinate is executed by the second spectroscope. Convolution and autocorrelation operations can be performed by similar devices.

References

1) A. J. Downes: Brit. Patent 129747 (1918). N. E. Lindenblad: U. S. Patent 2443258 (1948). A. I. Kartashev: Opt. Spectr. **9** (1960) 204.

2) M. Françon: Nuovo Cimento Suppl. **9** (1952) 283. W. Lukosz: Z. Naturf. **18a** (1963) 436. W. Lukosz and M. Marchand: Opt. Acta **10** (1963) 241. B. Morgenstern and D. P. Paris: J. Opt. Soc. Am. **54** (1964) in print.

3) A. Lohmann: Opt. Acta **3** (1956) 97. W. Gärtner and A. Lohmann: Z. Phys. **174** (1963) 18. A. Lohmann and D. P. Paris: Appl. Opt. **3** (1964) 1037.

4) E. Menzel: Naturwiss. **46** (1959) 105. E. Menzel and K. Hohlfeld: Z. Phys. **164** (1961) 522.

5) A. Lohmann and D. P. Paris: J. Opt. Soc. Am. **54** (1964) 579.

DISCUSSION

Nomarski, G. Il est bon de préciser c'est qui est la super- ou hyper-résolution. Dans le cas de filtrage en cohérence partielle proposé par moi et aussi dans le cas de l'apodisation optimale proposé par M. Lansraux la fréquence limite reste inférieure ou égale à la fréquence limite

$$f_M \leq \frac{2(N.\,A.)}{\lambda}\,.$$

Par contre, dans le cas de la une méthode de M. Lukosz on obtient (pour la 1e fois) l'information sur les fréquences spatiales $f>f_M$. Qu'en est il dans votre cas, M. Lohmann?

Lohmann, A. About the term "superresolution." If a given lens *with* additional devices (like Wollaston prisms, spectroscopes, moving masks) has a larger passband for spatial frequencies than this lens would have when used alone, then we talk about superresolution. In our case, it is the light pipe, which has zero bandwidth when used alone, but a finite bandwidth in connection with the 2 spectroscopes. For avoiding further controversies on names I might call this experiment simply "image transmission by lambda encoding."

Lansraux, G. I agree with Dr. Lohmann about the necessary conditions for superresolution or hyperresolution, i.e. the aberration of any conventional optical system, either in its use (for example scanning of the object and reconstruction of the image) or by adding parts of optical equipment in order to isolate a limited information from the object and to transfer it in the image with a considerably enhanced resolving power. Basically, according to the theory of optical transfer function, the frequencies transmitted by the instrument vanish when being higher than a finite limit. Conversely the effect of hyperresolution is to provide, in the reconstructed image, information which should correspond to a larger range of frequencies than the range transmitted by a conventional system. Coming back to my paper, presented previously, I would like to point out that an amplitude filter providing the equilibrium distribution $T(x)$ on the pupil corresponds obviously to a limited range of transmitted frequencies. Nevertheless this function is useful for hyperresolution because it is a steady distribution in an iterated diffraction process using diaphragms of radius W_m.

Ingelstam, E. I want to put your attention to the fact that we use the term information in a too narrow sense, namely, the sampling points given by resolution in a lens. Other information-carrying parameters in the (at least) four fold integral are: spectra, time, polarization. [Reference. Ingelstam: "On Problems in Contemporary Optics," Proc. of the Florence Meeting (1954).]

Lohmann, A. This is a comment on the relation between the superresolution and the classical theories of resolution due to *Abbe* and *Duffieux*. The following assumptions are made (but seldom mentioned) in classical resolution theory: (a) non-double-refracting objects (→scalar theory); (b) monochromatic illumination; (c) time-independent operation. It is not surprising that one can overcome the classical limit of resolution (=band-width for spatial frequencies), if one performs experiments which are not restricted by these three assumptions. For results see the references at the end of my paper.

Coleman, K. One should consider the light source as governing the total information content of an optical system. Any operation carried out on the light from a source is then of a sampling nature whether it is restriction by an aperture or dispersion by a prism or reception by a photographic plate. In the first case the light flux, as an information channel, includes some information about the boundary. In the second case, like Lohmann's experiment, the flux contains information concerning the dispersion system. Many approaches to information content of optical systems lack generality because they take apertures, objects or objectives as basis rather than operators or perturbations.

Reprinted from
Optica Acta, Vol. 12, No. 2, p. 185, April 1965

Rotary shearing interferometry†

by J. D. ARMITAGE, Jr. and A. LOHMANN

IBM Corporation
Monterey & Cottle Roads, San Jose, California, U.S.A.

(*Received* 26 *August* 1964)

The most common type of shearing interferometer produces two laterally shifted wavefronts. Longitudinal, radial and mirror reversed shifts have been investigated. We here consider shearing the two wavefronts rotationally to different azimuth angles [1]. This is particularly meaningful if the angular structure of the wavefront is of interest. An important example of such a wavefront is the aberrated wavefront of a lens. At different values of the shearing angle, different portions of the aberrations with specific degrees of rotational symmetry will appear or disappear. This allows one, for example, to concentrate on coma aberrations while spherical and astigmatic aberrations do not appear. The theory and several ways of realizing rotary shearing interferometry are described. Emphasis will be devoted to the solution of a coherence problem, which is specific for rotary shearing interferometry as pointed out by Murty [2].

1. Prehistory of the problem

Two-beam interferometers can be used for the observation of phase objects in two ways (compare figure 1): (1) either only one of the two beams hits the

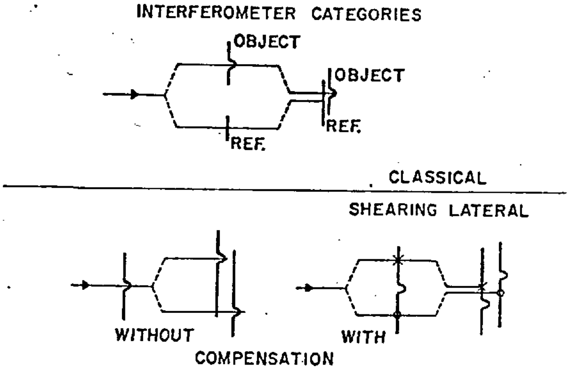


Figure 1. In a 'classical' interferometer (e.g. Michelson) the distorted object wave OBJ interferes with a plane reference wave REF. Lenses omitted; beam splitter indicated schematically. In a 'shearing' interferometer two identical distorted object waves interfere with each other. These two object waves are sheared, in this figure laterally. The contrast of the interference fringes will be improved by introducing 'compensation'. That is, each pair of rays, originating from the same illuminating ray, will be unified in the image plane.

† Presented at the ICO Interference and Coherence Conference, Sydney, Australia, August 1965 under the title *Theta Shearing Interferometry*.

object (typical examples are Michelson, Mach-Zehnder), (2) or both beams interact with the object, in which case the two images from the two beams are shifted with respect to each other (shearing interferometry). Lateral, longitudinal, and radial shifts are suitable, as is reversal of one of the two images. Here we will investigate azimuthal or rotary shearing. That is, one image is rotated with respect to the other one: $(r, \phi) \rightarrow (r, \phi+\theta)$. This type of shearing interferometry covers the same range of applications as do the others. Beyond that it is particularly useful if the angular aspect of the object is of special interest. This is the case in studying the wave aberrations of an optical system. For example, one can observe coma while suppressing spherical aberration and astigmatism. In other words, only object details with first order, rather than those of zero and second order, rotational symmetry will appear when θ is chosen appropriately.

The principle of rotary shearing interferometry and some of its virtues has been mentioned in 1947 by Bates [1]. A particular coherence problem, which arises when one of two interfering wavefronts is rotated, has been observed and analysed recently by Murty [2]. He used a Michelson interferometer for investigating the quality of reflecting phase objects such as roof prisms and corner cubes, where his objects themselves reversed or rotated by 180° one of the two temporarily separated wavefronts. He observed the interference between the object wave and a plane wave as reference. In this respect his instrument does not fall into the class of interferometers, where two sheared images of the same object interfere. However, the coherence problem of Murty is also pertinent to our type of shearing interferometer. As a consequence of this coherence problem, one gets good fringe contrast only in the centre of the image field. This requires using a very fine pinhole as source, sacrificing illuminance, or one has to illuminate by means of a single-mode laser.

In this paper we propose first how the rotary shearing principle can be applied for transparent phase objects, and second, we show schematically how this coherence problem can be overcome by incorporating a rotary shearing compensator. Finally we will describe some instruments which include such compensation.

2. An interference microscope for transparent objects, based on rotary shearing

The essential component in all our rotary shearing interferometers is an image reverser such as a Dove prism (in transmission) or a roof prism (in reflection). If the image reverser has the angular position α, measured around the optical axis, an image $u(r, 2\alpha-\phi)$ instead of $u(r, \phi)$ will appear. It might be mentioned that many other prisms can fulfil the same function and are sometimes more practical : Delaborne, Abbe-König, Schmidt-Pechan.

In our first proposal (figure 2 (*a*) and (*b*)) two such image reversers are incorporated with variable orientations α and β. They are located at places where the beam going from object to image plane is temporarily split into two beams. Hence from the object $u(r, \phi)$ two images will be created : $u(r, 2\alpha-\phi)+u(r, 2\beta-\phi)$. These images will interfere with each other if the light stems from a monochromatically illuminated pinhole. However, the contrast K will decrease rapidly with increasing radial coordinate r, radius ρ of pinhole, and shearing angle $\alpha-\beta$.

By slightly generalizing Murty's first formula one gets as contrast K:

$$K=\frac{2J_1(\nu)}{\nu},$$

where J_1 is the Bessel function of first order and first kind and

$$\nu=\frac{4\pi\rho r\sin(\alpha-\beta)}{\lambda f_1}.$$

The interferometer as shown in figure 2 (*a*) and figure 2 (*b*) can be replaced by a more stable and compact one (figure 3 (*a*)). Since only the difference of the angles α and β is of interest, one can afford to have one of the two roof prisms fixed.

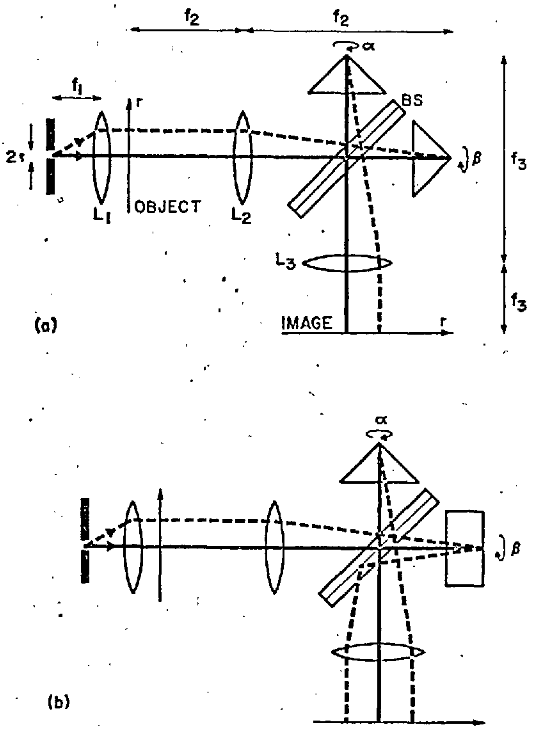


Figure 2. Interferometer which creates two azimuthally sheared images. The basic interferometer is of the Michelson type with two roof prisms as reflectors. The purpose might be to observe the wave aberrations of lens L_1. (*a*) Orientation of roof prisms $\alpha=0$, $\beta=0$, hence shearing angle $\theta=2\alpha-2\beta=0$. (*b*) $\alpha=0$, $\beta=90°$, $\theta=180°$.

Even better in terms of path length compensation is the Sagnac-type interferometer of figure 3 (*b*). Rotating the Dove prism will rotate the image in opposite directions, depending on which way the light passes through the square loop.

For good coherence the state of polarization of both beams must be identical. To achieve this one should include two polarizers, one before and one after the interferometer, both oriented at 0°. Furthermore one should include a ' polarization coupling ' [3] between the rotatable prism and the main prism. This coupling consists of two quarter-wave plates, one being fixed at 45° to the main prism, the other one at 45° to the hypotenuse of the rotatable β-prism. Between the two quarter-wave plates the light will be circularly polarized. However, within the roof prism it will be linearly polarized, always parallel to the roof ridge, so that phase jumps at internal reflections cannot alter the state of polarization. When returning into the main prism after passing again the two quarter-waveplates, the old direction of linear polarization is restored, independently of the angle β of the roof prism.

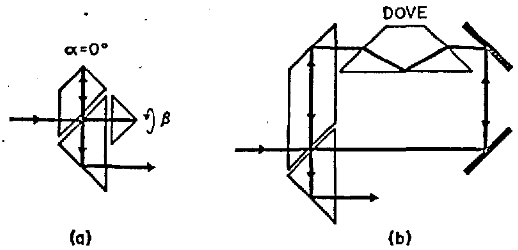


Figure 3. Two other interferometers, operating on the same principles as the interferometer of figure 1. (*a*) A more compact version of the Michelson type; one roof prism fixed. (*b*) Sagnac type interferometer; shearing angle is equal to four times the Dove rotation; Dove prism should be in rear focal plane of L_2 (see figure 1) and in front focal plane of L_3, for both beam directions within the loop.

3. Principle of rotary shearing compensation

To better understand Murty's coherence problem, let us remember a sufficient condition for good spatial coherence of two-beam shearing interference in general. When looking through the whole apparatus at the source, the two images of the source should coincide. That excludes not only a relative shift or tilt, but also a rotational shear of the two source images, as one would observe in Murty's experiment. However, at the same time we do *not* want the two images of the object under investigation to be coincident, since they are to be sheared. One can overcome this problem by splitting up the original rays *before* they hit the object. Each pair of rays then experiences a rotary shear before hitting the object, just enough to compensate the main shear, which takes place after the object. The two shearing operations then cancel each other as far as coherence is concerned, but only the second shear acts upon the light after it has interacted with the object. Hence two images of the object rotated with respect to each other will be observed.

A more quantitative and schematic explanation of this shearing compensation can be gained by reference to figure 4. Perfect coherence is achieved if the difference in shearing angle of a pair of illuminating beams is compensated to zero when arriving at the image plane. A net shearing angle $\theta = 2\gamma - 2\delta$ of the two images remains. There are essentially two particular solutions of the coherence equation, whereby one can save two out of four Dove prisms:

$$\beta = 0 = \delta; \quad \alpha = \gamma$$
$$\alpha = 0 = \delta; \quad \beta = -\gamma.$$

Two plane parallel glass plates might replace the two Dove prisms with fixed orientation at zero degrees in order to compensate for path differences which otherwise would destroy the coherence. However, in that case, the two images would be not only rotated with respect to each other, but also reversed. It should be noted that an analogous compensation scheme for lateral shearing interferometry has been used already by many authors.

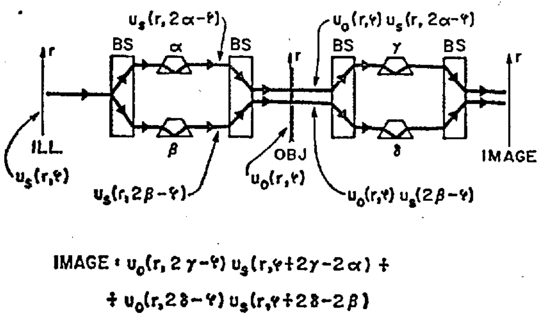

$u_s(r, 2\alpha-\phi)$
$u_0(r,\phi)\, u_s(r, 2\alpha-\phi)$
BS
α
BS
BS
γ
BS
r
r
r
ILL.
β
OBJ
δ
IMAGE
$u_s(r,\phi)$
$u_s(r, 2\beta-\phi)$
$u_0(r,\phi)$
$u_0(r,\phi)\, u_s(2\beta-\phi)$
IMAGE : $u_0(r, 2\gamma-\phi)\, u_s(r, \phi+2\gamma-2\alpha)$ +
+ $u_0(r, 2\delta-\phi)\, u_s(r, \phi+2\delta-2\beta)$

Figure 4. Scheme of a general rotary shearing interferometer. Two Dove prisms before object, two others behind object; BS = beam splitter; lenses omitted. A Dove prism oriented at α changes $\phi-\alpha$ into $-(\phi-\alpha)$ or ϕ into $2\alpha-\phi$.

4. Compensated rotary shearing interferometers

We shall now describe some rotary shearing interferometers which incorporate the shearing compensation principle, permitting an extended source to be used without sacrificing contrast.

The set-up of figure 5 (*a*) is the most direct implementation of the schematic set-up of figure 4. Four Wollaston prisms W_1 . . . W_4 act as beam splitters. Since for this type of beam splitter it is essential to preserve the state of polarization, one has to surround the rotatable Dove prisms by ‘ polarization couplers ’ as mentioned previously. In this case the coupling can be accomplished by putting four quarter-wave plates at 45° orientation before W_2 and W_4 and after W_1 and W_3. Furthermore, two quarter-wave plates should be attached to both ends of each Dove prism with a 45° relative angle.

The arrangement of figure 5 (*a*) can be simplified considerably (figure 5 (*b*)). Not only can two Wollaston prisms and three lenses be saved, but also ‘ polarization coupling ’ will be much easier. For example, the rotary shearing might be accomplished by setting the Dove prisms as follows:

$$\text{Dove 1} : \alpha, \quad \text{Dove 4} : -\alpha, \quad \text{Dove 2} : 0^\circ, \quad \text{Dove 3} : 0^\circ.$$

The two prisms Dove 2 and Dove 3 might be taken out and be replaced by thick glass plates for path length compensation. The problem of polarization coupling can be solved by using, instead of many quarter-wave plates, a single half-wave plate inserted between Dove 1 and Dove 4 and oriented parallel to the wedge of the Wollaston prisms. This half-wave plate rotates the plane of polarization of that portion of the beam which is polarized *parallel* to the main plane of Dove 1, so that it will be *perpendicular* when interacting with Dove 4. Since both portions of the Dove 1–Dove 4 beam will experience in total the same amount of phase jump at refractions and reflections within the two Dove prisms, the state of polarization will be linear again and properly oriented when arriving at the final Wollaston prism. This simplified polarization coupling is similar to the one proposed by Drougard and Wilczynski [4].

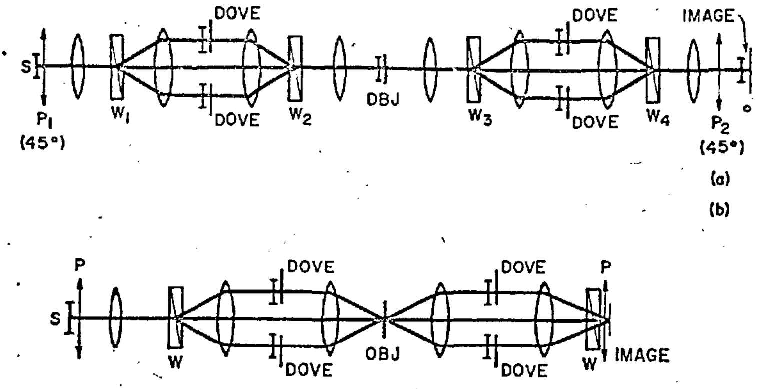


Figure 5. Compensated rotary shearing interferometer with Wollaston prisms as beam splitters. The Dove prisms are indicated by a straight line. (*a*) Four Wollaston prisms and at least eight quarter-wave plates are needed. (*b*) Only two Wollaston prisms and one half-wave plate are needed.

In figure 6 we show a device which is based on two beam-splitting units as shown earlier in figure 3 (*a*), essentially two compact Michelsons each with two roof prisms *R*. The rotatable prisms R_2 and R_4 should be set at α and $-\alpha$ to achieve shearing compensation. To both R_2 and R_4 a polarization coupler should be attached as described in connection with figure 3 (*a*).

Two beam splitters in sequence create four rays out of one original ray. For an easy image assessment it is desirable to have only two out of the four rays interacting in the image plane. This selection can be achieved by applying an idea of Hariharan and Sen [5]. For this we have to include means to rotate the plane of polarization by 90° when reflected at R_1 or R_3. Then the analyser P_2 can be used to eliminate two out of four rays. This rotation of polarization at R_1 and R_3 can be achieved by splitting up the roof into two equal parts and then inserting a half-wave plate of proper orientation between the parts.

Sagnac interferometers have the advantage of being automatically path length compensated, because the pair of the two beams travels through the same closed loop in opposite directions. We insert a phase object and a Dove

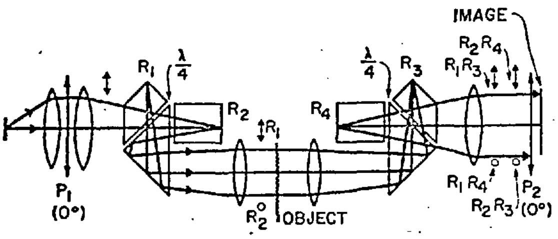


Figure 6. Compensated rotary shearing interferometer with Michelson type beam splitter. Roof prisms R_1 and R_3 fixed at zero-degree orientation; the other two, R_2 and R_4, rotatable, shown here at 90° orientation. Two quarter-wave plates at R_2 and R_4 in connection with two polarizers P_1 and P_2 allow selection of two out of four beams: R_1R_3 and R_2R_4, or R_1R_4 and R_2R_3. A half-wave plate near the object acts as phase jump compensator similar to that shown in figure 4 (*b*).

prism into the loop in figure 7. The sequence of passing the object and the Dove prism is of course opposite for the two oppositely travelling beams. The only image of the object which will be rotated when the Dove prism is rotated is the one for which the light has passed first the object and then the Dove prism. However, the other light beam will experience also a rotation before passing the object. This is just what is needed for compensation in order to get good interference contrast. The rotated image is also reversed:

$$u(r,\phi)+u(r,2\alpha-\phi).$$

By comparing figures 7 and 3 (*b*), one sees that the Sagnac loop may have different shapes: triangle or square. This is done purposely. One has to consider whether the number of reversions due to reflections is even or odd. Also one has to take into account the fact that in one case the object is within the loop, in the other case, before it.

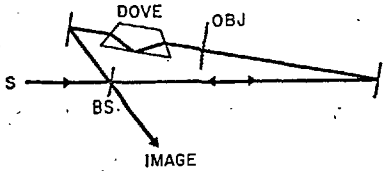


Figure 7. Compensated rotary shearing interferometer of the Sagnac type. Lenses and polarization coupling omitted.

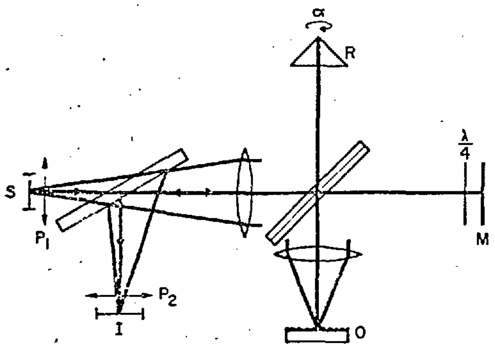


Figure 8. Compensated rotary shearing interferometer of the Hariharan and Sen type. Polarization coupling omitted.

So far we have proposed only methods applicable to transparent phase objects. Now we will show how the interferometer of Hariharan and Sen [5] can be modified to produce rotary shearing interference for a reflecting object O (figure 8). The Hariharan and Sen interferometer differs from the Michelson in that only those four beams are used which interact four times (rather than twice) with the beam splitter:

SMOMI, SMORI, SROMI, SRORI.

The polarizers P_1 and P_2 in connection with the quarter-wave plate in front of the mirror M are needed to implement the special idea of Hariharan and Sen which was mentioned earlier. It enables one, by setting the two polarizers either parallel or perpendicular, to select two out of the four rays : MOM and ROR, or MOR and ROM. The rotatable roof prism should be provided with polarization coupling consisting of two quarter-wave plates as described before. This interferometer is shear compensated. Its two images are sheared by 2α and also reversed as in the Sagnac version.

Les interféromètres les plus courants à dédoublement après traversée de l'objet produisent deux fronts d'onde décalagés latéralement. Les décalages longitudinal, radial et par renversement de l'une des deux images ont été étudiés déjà. Nous envisageons ici le dédoublement des deux fronts d'onde par rotation suivant des azimuths différents [1]. Ceci est particulièrement utile si on s'intéresse à la structure angulaire du front d'onde. Un exemple important d'un tel front d'onde est fourni par la surface d'onde aberrante d'une lentille. Pour diverses valeurs de l'angle de décalage théta, différentes portions des aberrations avec des degrés caractéristiques de symétrie de révolution apparaîtront ou disparaîtront. Ceci permet, par exemple, d'étudier uniquement les aberrations de coma, tandis que l'aberration sphérique et l'astigmatisme n'interviennent pas. On décrit la théorie et quelques façons de réaliser des interféromètres à dédoublement par rotation. Nous insisterons sur la solution d'un problème de cohérence, qui est spécifique à l'interférométrie par dédoublement par rotation, ainsi que l'a remarqué Murty [2].

Die meist benutzte Bauart des Shearing-Interferometers erzeugt zwei seitlich versetzte Wellenfronten. Es sind aber auch schon longitudinale, radiale und spiegelverkehrte Versetzungen untersucht worden. Wir betrachten hier ein Shearing an zwei Wellenfronten, die auf verschiedene Beträge des Azimutwinkels verdreht sind [1]. Das hat namentlich dann eine Bedeutung, wenn die Winkelstruktur der Wellenfront untersucht werden soll. Ein wichtiges Beispiel dafür ist die mit Aberrationen behaftete Wellenfront einer Linse. Bei verschiedenen Werten des Shearingwinkels ν erscheinen oder verschwinden verschiedene Anteile der Aberrationen mit einem spezifischen Wert der Rotationssymmetrie. Dies erlaubt z.B. die Koma herauszuheben, während die sphärischen und astigmatischen Abweichungen unterdrückt werden. Die Theorie und verschiedene Wege zur Ausführung eines Rotations-Shearing-Interferometers werden dargestellt. Besonderer Nachdruck wird auf die Lösung eines Kohärenzproblems gelegt, das für die Rotations-Shearing-Interferometrie wesentlich ist, wie Murty [2] festgestellt hat.

References

[1] Bates, W. J., 1947, *Proc. phys. Soc., Lond.*, **59**, 940.
[2] Murty, M. V. R. K., 1964, *J. opt. Soc. Amer.*, **54**, 571.
[3] Steel, W. H., 1964, *Opt. Acta*, **11**, 9.
[4] Drougard, R., and Wilczynski, J. S., 1964, *J. opt. Soc. Amer.*, **54**, 1406 A.
[5] Hariharan, P., and Sen, D., 1961, *Proc. phys. Soc., Lond.*, **77**, 328.

Reprinted from APPLIED OPTICS, Vol. 4, page 461, April 1965

Character Recognition by Incoherent Spatial Filtering

J. D. Armitage and A. W. Lohmann

The character recognition method described here is based on the principle of incoherent spatial matched filtering. The input to this matched filter is not the unknown character itself, but its Fraunhofer diffraction pattern. The intensity distribution in this diffraction pattern is insensitive against shifting of the unknown character, avoiding the need for character registration. The incoherent matched filter is easier to implement than the coherent matched filter, since only binary rather than continuous-tone masks are required. The theory and some experiments will be discussed and compared with other optical character recognition methods.

I. Introduction

It remains yet to be seen to what extent optical methods will ultimately contribute to the general field of data processing. However, an optical approach has undeniable attraction in those areas where the data are inherently two dimensional. It has been pointed out (for example, by Horwitz and Shelton[1] in 1961) that application of Fourier optical processing should apply particularly well to optical character recognition, solving additionally the character registration problem simultaneously with the recognition problem. Horwitz and Shelton pointed out three principal difficulties: (a) lack of light, especially for multichannel processing, (b) the need for transparency input, and (c) the need for continuous-tone photographic masks which may be difficult to prepare. Recent developments, such as the laser, widespread application of microfilm techniques, and devices using the Eidophor[2] process help to alleviate the first two of these problems. The method described here circumvents the third problem.

The recognition method is based on spatial matched filtering for incoherent objects. General spatial filtering has been applied more often for the coherent[3] than for the incoherent[4] case. More recently, spatial filtering has been extended to complex matched filtering for the coherent case in important work by Leith and Upatnieks,[5] and Vander Lugt.[6] Kelly,[7] and Trabka and Roetling[8] have described an approach to incoherent matched filtering based on geometric-optical shadowing. We here present a method of matched filtering which is rigorous in terms of the diffraction theory of incoherent image formation.

In Sec. II is described the basic concept of matched filtering as it applies to the character recognition problem. In Sec. III is discussed the mathematical equivalence and experimental differences between a spatial matched filter and a correlation operation. In Sec. IV the preprocessing method of Horwitz and Shelton, is explained, which results in a signal whose position is independent of the character position. Section V describes the synthesis of a set of matched filters to permit recognition of the unknown character by further processing this signal. In Sec. VI a method of masking applied to the input plane of the matched filter system to improve the recognition discrimination is dealt with. In Sec. VII a new method of parallel processing to permit an increase in processing speed is present. In Sec. VIII the experimental implementation of this matched filter method is reported. Finally, in Sec. IX, some of the more practical differences between our method and other methods of optical character recognition are emphasized.

II. Pattern Recognition by Spatial Matched Filtering

The well-known concept of matched filtering serves two purposes in electrical engineering: detection and recognition of signals. The basic concept has been translated into optics and used for the detection of spatial patterns by Kelly, Trabka and Roetling, and by Vander Lugt, as mentioned in Sec. I. These authors wished to find the unknown position of a signal which is surrounded by noise. The matched filter solves this problem by maximizing the ratio of peak signal energy to mean-square (white) noise energy in the output of the processing system.

The authors are with IBM Corporation, San Jose, California.
Received 12 December 1964.

This paper, entitled "Shift Invariant Optical Pattern Recognition," was presented at the October 1964 meeting of the Optical Society of America.

Our goal is to recognize an unknown input or optical pattern $I_n(x)$ at a known position, say around $x = 0$.[*] The object $I_n(x)$ is only partially unknown, in that it is known to be a member of the set or font $I_n(x)(n = 1, 2, \ldots, N)$. The pattern $I_n(x)$ is not the unknown character itself. but is derived from it by means of a preprocessing operation as described in Sec. IV.

Recognition of the input pattern may be achieved by the following experimental operations: (a) an image of the input $I_n(x)$ is generated by an image-forming system with an optical transfer function (OTF) $T_m(\nu)$; (b) a pinhole with a photodetector behind it is placed at the center (on the optical axis) of this image plane; (c) the "correlation coefficients" S_{nm} $(m = 1, 2, \ldots, n, \ldots, N)$ are sequentially measured as the photocell signal by modifying the OTF of the system, $T_m(\nu)(m = 1, 2, \ldots, n, \ldots, N)$ to N different states.

The problem is now to find a set of N OTF's $T_m(\nu)$, or spatial frequency filters, such that the unknown input $I_n(x)$ can be recognized by appropriate comparison of the correlation coefficients, S_{nm}. It is convenient to choose a set of OTF's $T_m(\nu)$ such that the autocorrelation coefficient S_{nn} will be larger than the competing crosscorrelation coefficients S_{nm} $(m \neq n)$ or $S_{nm}/S_{nn} < 1$ $(m; n = 1, 2, \ldots, N;$ but $m \neq n)$.

A calculation of the correlation coefficients S_{nm} follows; the input may be written as a Fourier integral

$$I_n(x) = \int \tilde{I}_n(\nu) \exp(2\pi i\nu x)d\nu.$$

The OFT $T_m(\nu)$ modifies the spatial spectrum $\tilde{I}_n(\nu)$ of the object, generating an image

$$I_{nm}(x) = \int \tilde{I}_n(\nu)T_m(\nu) \exp(2\pi i\nu x)d\nu.$$

In the pinhole at $x = 0$ one gets

$$I_{nm}(0) = S_{nm} = \int \tilde{I}_n(\nu)T_m(\nu)d\nu,$$

or, in abbreviated form, $S_{nm} = \{\tilde{I}_n T_m\}$.

Our problem can be solved by choosing as OTF's the well-known set of matched filters $T_m(\nu) = \tilde{I}_m{}^*(\nu)/S_{mm}$ $(m = 1, 2, \ldots, N)$. Each of these filters is proportional to the complex conjugate of the spatial spectrum of one of the possible inputs. The proof that these matched filters satisfy the recognition inequality $S_{nm} < S_{nn}$ $(m = 1, 2, \ldots, N)$ rests on the Schwartz[†] inequality:

$$\left(\frac{S_{nm}}{S_{nn}}\right)^2 = \frac{\{\tilde{I}_n \tilde{I}_m{}^*\}^2}{\{\tilde{I}_n \tilde{I}_n{}^*\}\{\tilde{I}_m \tilde{I}_m{}^*\}} < 1.$$

This inequality holds for all cases where $\tilde{I}_n(\nu)$ and $\tilde{I}_m(\nu)$ are significantly different: $\tilde{I}_m(\nu) \neq c\tilde{I}_n(\nu)$ (for any arbitrary constant factor c). In other words, any wrong guess about $I_n(x)$ with a filter $I_m(\nu)(m \neq n)$ yields a correlation coefficient S_{nm} smaller than the correct correlation coefficient S_{nn} which would have resulted from filter $T_n(\nu)$.

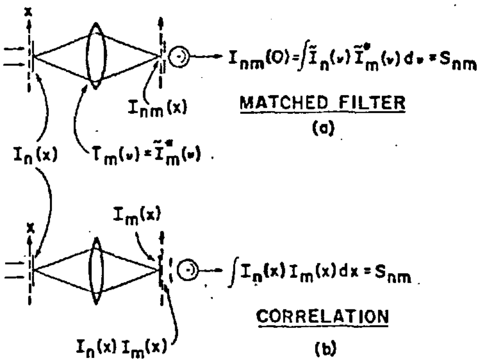

$I_{nm}(0) = \int \tilde{I}_n(\nu)\tilde{I}_m^*(\nu)d\nu = S_{nm}$
$I_{nm}(x)$
MATCHED FILTER
(a)
$I_n(x)$
$T_m(\nu) = \tilde{I}_m^*(\nu)$
$I_m(x)$
$\int I_n(x) I_m(x) dx = S_{nm}$
CORRELATION
$I_n(x) I_m(x)$
(b)

Fig. 1. Optical systems for (a) matched filter and (b) correlation operations.

III. Comparison between Matched Spatial Filters and Spatial Correlation

It is well known that the two operations, matched filtering and correlation, are mathematically equivalent, but sometimes quite different in terms of implementation. We will now investigate this difference for the case of optical filtering and correlation, in order to enable us to compare later the matched filter method with Horwitz and Shelton's correlation experiment. Parseval's theorem can be written as

$$\{\tilde{I}_n \tilde{I}_m{}^*\} = \{I_n I_m{}^*\}, \text{ or } \int \tilde{I}_n(\nu)\tilde{I}_m{}^*(\nu)d\nu = \int I_n(x)I_m{}^*(x)dx.$$

It follows from this equality that two different optical experiments, matched filtering and correlation (Fig. 1), will always yield the same results. In Fig. 1(a) illustrating matched filtering, the object is the incoherently illuminated transparency $I_n(x)$. The lens has a transfer function $T_m(\nu)$ which is one of the set of matched filters, $T_m(\nu) = \tilde{I}_m{}^*(\nu)$, neglecting now the constant factor $1/S_{mm}$. A pinhole is placed in the center of the image plane, and a photocell behind it measures the correlation coefficient S_{nm}. This same correlation coefficient can also be measured by the experiment shown in Fig. 1(b) illustrating correlation. Here the optical system forms a perfect image [$T(\nu) = 1$ for the frequency range of significance] of the unknown pattern $I_n(x)$, and a photocell measures the total flux passing through the known pattern $I_m(x)$. Both experiments, although mathematically equivalent, may be quite different in terms of implementation. Which will be preferred must be decided for each specific case, taking into account the type of mask (or filter) and photodetector required for each experiment.

IV. Preprocessing for Shift Invariance

It is well-known that a lens is able to perform a Fourier transform in two dimensions. More precisely, a monochromatically and spatial-coherently illuminated amplitude transmission $u_n(x_0)$ in the front focal plane of the lens will give rise to a complex amplitude $a_n(x)$ in the rear focal plane such that

[*] One-dimensional notation is used for convenience only, although the patterns are two dimensional.

[†] See, for example, H. Margenau and G. M. Murphy, *The Mathematics of Physics and Chemistry* (Van Nostrand, New York, 1947), p. 131.

$$a_n(x) = \tilde{u}_n(x/\lambda f);\ \tilde{u}_n(\nu) = \int u_n(x_0)\exp(-2\pi i x_0 \nu)dx_o.$$

It follows from a known property of Fourier transformation that the intensity $I_n(x)$ in the rear focal plane will not be altered if the amplitude object $u_n(x_0)$ is shifted laterally: if $u_n(x_0) \rightarrow u_n(x_0 + x_n)$, then $a_n(x) \rightarrow a_n(x)\ \exp[i\phi(x,x_n)]$, where $\phi = 2\pi x x_n/\lambda f$; the intensity $I_n(x)$ equals $|a_n(x)|^2$, which is independent of ϕ and, hence, of x_n. The intensity distribution, $I_n(x)$, is also invariant against shift of the amplitude object along the optical axis, if the object is illuminated by a plane wave. This can be shown using the equation describing the propagation of plane waves behind an object,[9]

$$u(x,z) = \int \tilde{u}_0(\nu)\exp[2\pi i(\nu x + \sqrt{1 - \lambda^2\nu^2}\, z/\lambda)]d\nu,$$

where

$$\tilde{u}_0(\nu) = \int u(x,0)\exp(-2\pi i\nu x)dx;$$

if $u_n(x_0,0) \rightarrow u_n(x_0,z_n)$, then $a_n(x) \rightarrow a_n(x)\exp[i\psi(x, z_n)]$, where $\psi = 2\pi\sqrt{1- (x/f)^2}\, z_n/\lambda$, so $I_n(x) = a|_n(x)|^2$; or it can be considered as a special case of Toraldo di Francia's principle of "inverse interference".[11]

This means that, for our character recognition experiment, the unknown primary signal amplitude $u_n(x_0)$ is transformed by the preprocessor into the secondary signal intensity $I_n(x)$, such that the secondary signal is not affected by any lateral or longitudinal shift of the primary signal. The secondary signal acts as input for the matched filter system described in Sect. II. We must be certain that any phase relationships in the rear focal plane of the preprocessing transform lens are destroyed, because the phase would *remember* the position of the primary signal. This is achieved by inserting a rotating ground glass, which eliminates this phase information in the time average.

In Fig. 2, the basic setup is shown. The laser illuminates coherently the unknown character $u_n(x_0)$, and the preprocessing lens transforms $u_n(x_0)$ into the intensity $I_n(x)$ just after the ground glass. The incoherent matched filter, $T_m(\nu)$, which follows, has as its main constituent the pupil function $p_m(x')$. This matched filter transforms its input $I_n(x)$ into the output $I_{nm}(x)$.

V. Synthesis of the Incoherent Matched Filter

The problem now is to achieve the matched filter functions $T_m(\nu) = \tilde{I}_m{}^*/S_{mm}(m = 1,2,\ldots,N)$ by choosing the proper pupil functions $p_m(x')$ $(m = 1,2,\ldots,N)$ which are to be implemented as complex transmission factors in the image-forming system of Fig. 2. The problem would be an easy one, at least in principle, if we dealt with coherent matched filtering, where the pupil function itself constitutes the filter function.

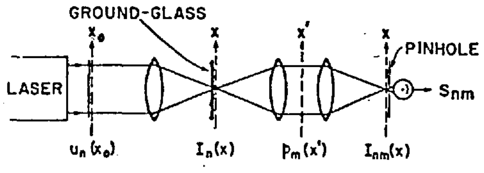


Fig. 2. Optical system for character recognition, including preprocessing for shift invariance and incoherent matched filtering.

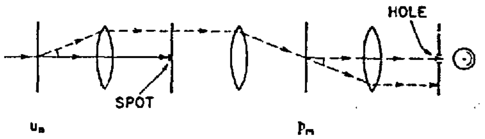


Fig. 3. Only light which has been twice diffracted at the same angle, by the unknown character and by the filter, will pass through the hole to the photodetector.

However, here the object $I_n(x)$ is incoherently illuminated due to the rotating ground glass having randomized the phase relationships in the complex amplitude. Therefore, the relationship between filter function (or OTF) and pupil function is less direct, as has been derived by Duffieux[11]:*

$$T_m(\nu) = \int p_m\left(x' + \frac{\lambda f\nu}{2}\right) p_m{}^*\left(x' - \frac{\lambda f\nu}{2}\right)dx'.$$

In general, it is not an easy problem to prescribe first the OTF and then look for a suitable pupil function p_m. However, in the present case, the matched filter condition, $T_m(\nu) = \tilde{I}_m{}^*(\nu)/S_{mm}$, can be satisfied fairly easily, as follows. Our input, $I_n(x)$, to the matched filtering operation was generated from the original character in front of the transform lens. By definition,

$$\tilde{I}_m{}^*(\nu) = \int I_m(x)\exp(2\pi i\nu x)dx,$$

where

$$I_m(x) = |a_m(x)|^2 = |\tilde{u}_m(x/\lambda f)|^2,$$

$$\tilde{u}_m(\nu) = \int u_m(x_0)\exp(-2\pi i\nu x_0)dx_0.$$

Substituting these values leads to

$$\tilde{I}_m{}^*(\nu) = \int u_m\left(x_0 + \Delta x + \frac{\lambda f\nu}{2}\right) u_m{}^*\left(x_0 + \Delta x - \frac{\lambda f\nu}{2}\right)dx_0,$$

with arbitrary shift Δx. Now the condition for matched filtering reduces to:

$$T_m(\nu) = \int p_m\left(x' + \frac{\lambda f\nu}{2}\right) p_m{}^*\left(x' - \frac{\lambda f\nu}{2}\right) dx'$$
$$= \left[\int u_m\left(x_0 + \Delta x + \frac{\lambda f\nu}{2}\right) u_m{}^*\left(x_0 + \Delta x - \frac{\lambda f\nu}{2}\right) dx_0\right]\frac{1}{S_{mm}}.$$

A simple solution offers itself immediately:

$$p_m(x') = u_m(x' + \Delta x)/\sqrt{S_{mm}}.$$

This is not necessarily the only solution, but it is a very practical one. It means the pupil functions $p_m(x')$ should have exactly the same complex amplitude transmissions as the set of possible unknown original charac-

* Note: If one thinks of x' and ν as two-dimensional vectors, this Duffieux formula holds also for the two-dimensional case. It should be noted that here we do not normalize to $T(0) = 1$ as is usually done, since here the ratio of mean radiance in both object and image planes is of interest.

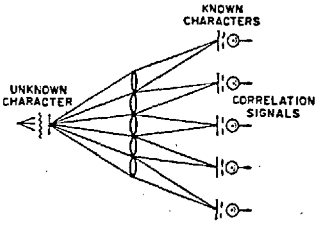


Fig. 4. Conventional system extended to parallel processing.

ters $u_m(x_0)$, and that they may be placed at an arbitrary position Δx in the filter plane. Also, the constant facfor $1/\sqrt{S_{mm}}$ must be included, which can be implemented easily either as an absorbing filter or electronically. The production of the proper pupil functions $p_m(x')$ becomes particularly simple if the unknown original characters $u_m(x_0)$ contain only opaque and fully transparent areas.

This simple result–that the proper pupil functions, $p_m(x)$, and unknown characters, $u_m(x)$–are identical, should not be considered as trivial. Since a ground glass is inserted between $u_n(x)$ and $p_m(x')$ to achieve insensitivity against shift of the original character, there is no image of the unknown character falling onto the pupil function, as in the case of the most elementary optical character recognition method [Fig. 1(b)]. On the contrary, the ground glass has to be sufficiently fine that the illumination falling onto the pupil function at any one instant does not show any structure.

This solution is rigorous in the realm of diffraction theory, whereas Kelly's and Trabka and Roetling's realization of a matched filter (which could be called shadow correlation) is based on geometric-optical considerations. This means, in practice, that we can fully utilize the image-forming qualities of the optical system, and do not reduce the resolution by strong defocusing.

VI. Improvement of the Discrimination Ratio

We have shown how to realize a set of matched filters $T_m(\nu)$ such that recognition is based on the inequality of correlation coefficients, $S_{nn} > S_{nm}$ $(m \neq n)$. However, sometimes the largest S_{nm} $(m \neq n)$ might be very nearly equal to S_{nn} for example, if the character Q is compared to an O. We will call $S_{nn}/(S_{nm})_{\max}$ the discrimination ratio which we want to increase.

To understand this approach qualitatively we recollect that diffraction takes place twice in our setup (Fig. 3). Let us make a distinction between *diffracted light* and *direct light*, the first deviated when interacting with u_n or p_m, the latter not deviated. The nature of the direct light is determined primarily by the size of the diffracting character, whereas the nature of the diffracted light is determined also by the structure of u_n or p_m. Since the size of the characters, or pupil masks, is irrelevant, we will prevent any direct light from reaching the photocell. This can be achieved by inserting an opaque spot into the ground glass plane (Fig. 3). The image of the opaque spot, formed by the optical system without any filter in place, would cover the pinhole.

Intuitively, one feels that this procedure should improve the recognition discrimination. We will now show quantitatively that this modified setup with a central spot [or any other transparency $T_s(x)$] in the ground glass plane still allows recognition by choice of the largest correlation coefficient $S_{nn} > S_{nm}$ $(m \neq n)$. The input of the matched filter system now is $I_n(x)T_s(x)$ instead of $I_n(x)$. The influence of the matched filter $T_m(\nu)$ is taken into account by multiplying the modified input spectrum,

$$\int I_n(x)T_s(x)\ \exp(-2\pi i\nu x)dx,$$

by

$$T_m(\nu) = \frac{\bar{I}_m{}^*}{S_{mm}} = \frac{1}{S_{mm}} \int I_m(x)\ \exp(2\pi i\nu x)dx$$

to get the spatial spectrum at the output plane. A Fourier transform yields $I_{nm}(x)$, the intensity distribution in the image plane. The correlation coefficient S_{nm} is then:

$$S_{nm} = I_{nm}(0) = \frac{1}{S_{mm}} \int I_n(x)T_s(x)I_m(x)dx.$$

From this, it follows that

$$S_{mm}{}^2 = \int I_m{}^2(x)T_s(x)dx$$

and

$$\frac{S_{nm}}{S_{nn}} = \frac{\int I_n(x)T_s(x)I_m(x)dx}{\sqrt{\int I_m{}^2T_s dx}\ \ \sqrt{\int I_n{}^2T_s dx}}.$$

If we substitute $\hat{I}_n(x) = I_n(x)\sqrt{T_s(x)}$, this expression will assume again the form of the Schwartz inequality, and hence the desired result that $S_{nn} > S_{nm}$ will be achieved.

The mask, $T_s(x)$, which we have introduced into the input plane of the matched filter, is simple to implement and is effective in improving the discrimination ratio, as described in Sec. VIII, but since this mask is based on an intuitive argument, it is not necessarily the theoretical optimum.

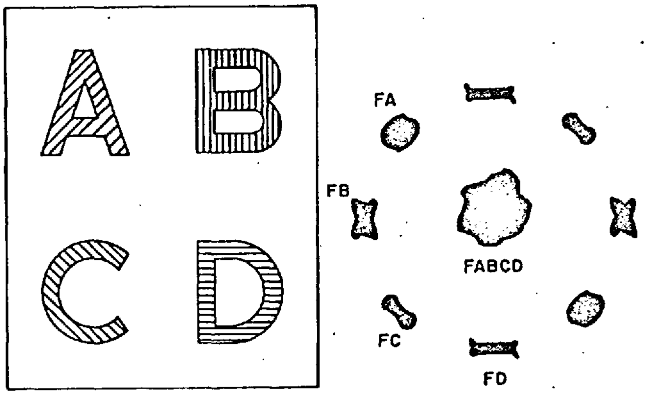


Fig. 5. Parallel processing (four patterns) using theta-modulated filters. At the left are the four filters theta-modulated with four different values of theta; at the right are (schematically) the output spectra. The unknown character is an F.

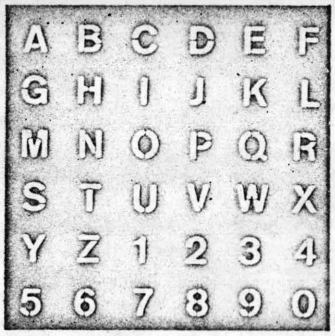


Fig. 6. The stencil-like photoetched characters used in the experiments. Each character is 1 mm high.

VII. Parallel Processing

So far we have considered only serial processing or sequential interaction of the unknown character with all possible matched filters. To improve the processing speed, it would be desirable to measure simultaneously all correlation coefficients originated by the unknown character. The operation of parallel processing can be introduced without any influence on the recognition principle. To use a set of lenses in parallel is one obvious solution, as is shown in Fig. 4 in connection with recognition by *direct correlation.*

Another way to implement parallel processing would be to set more than one pupil function $p_m(x')$ (such as $ABCD$) side by side in the matched filter system and superimpose gratings of different orientation onto the letters (left-hand side of Fig. 5). These gratings of different orientations will deviate the light into different directions, forming several diffraction patterns in the image plane (right-hand side of Fig. 5). Pinholes at the centers of these patterns, with photocells behind them, would allow measurement of several correlation coefficients simultaneously. The pattern on the optical axis gets light from all pupil functions $ABCD$, and, hence, this light, which is not useful for the recognition procedure, is wasted. Gratings without zero-order diffraction, or prisms, would avoid this loss. This method of parallel processing is based upon the principle of theta modulation.[12]

VIII. Experiment

We have performed some experiments to verify the theory developed in the preceding sections and to determine some of the practical difficulties which might be encountered in a physical application of these ideas. The system shown in Fig. 2 was set up on an optical bench, using a basic focal length of 200 mm. Our characters were about 1 mm size, giving a Fraunhofer diffraction pattern of about $\lambda f/d = 0.1$ mm without requiring an excessively long bench (<3 m). For convenience, we used a standard He–Ne laser, operating at 6 mW single-mode output. Our spot was made of photoopaquing paint of 120-μ diam on a thin plane-parallel glass plate. The diffuser must be negligibly thin in the z direction in order not to spread the Fraunhofer spectum falling upon it, have a structure size small relative to that of the diffraction pattern falling upon it in order to completely destroy the phase part of the spectrum, and be a random diffuser to obtain complete shift invariance. We have found that Scotch brand Magic Mending tape, rotated during the measurement procedure, works well.

An experimental problem stems from the requirement that the output correlation patterns be measured exactly at the ($x = 0$, $y = 0$) point in the pattern. It would be possible, although very inefficient, to measure the value of the function in (x,y) and determine its center from its symmetry, but this is not attractive even for a laboratory experiment. The alternative is to align the system by maximizing the output signal with circular filters (whose Airy patterns are known to have their maximum at $x = 0$, $y = 0$) rather than complicated characters, and then be certain that insertion of a character does not disturb this alignment. If the characters are on photographic emulsion, this will not usually be the case because of the gradients in backing thickness.[13] A liquid immersion cell[14] could eliminate this problem (and, additionally, eliminate phase effects due to emulsion reliefing) and is a practical solution, but we chose instead to photoetch the patterns through

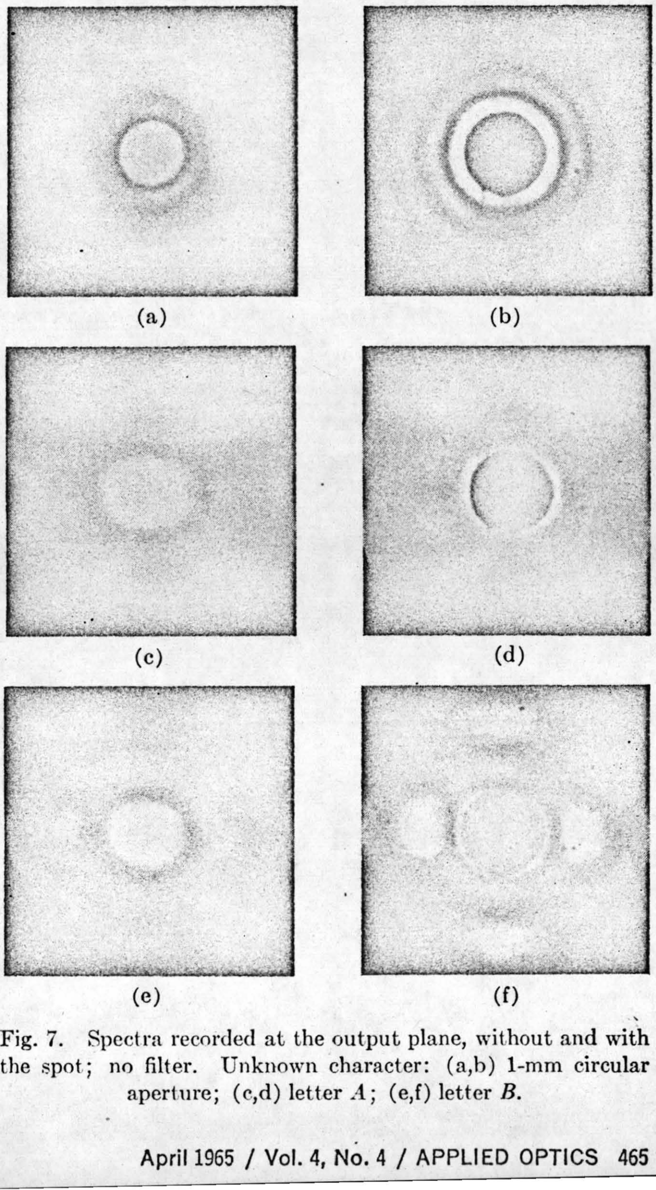


Fig. 7. Spectra recorded at the output plane, without and with the spot; no filter. Unknown character: (a,b) 1-mm circular aperture; (c,d) letter A; (e,f) letter B.

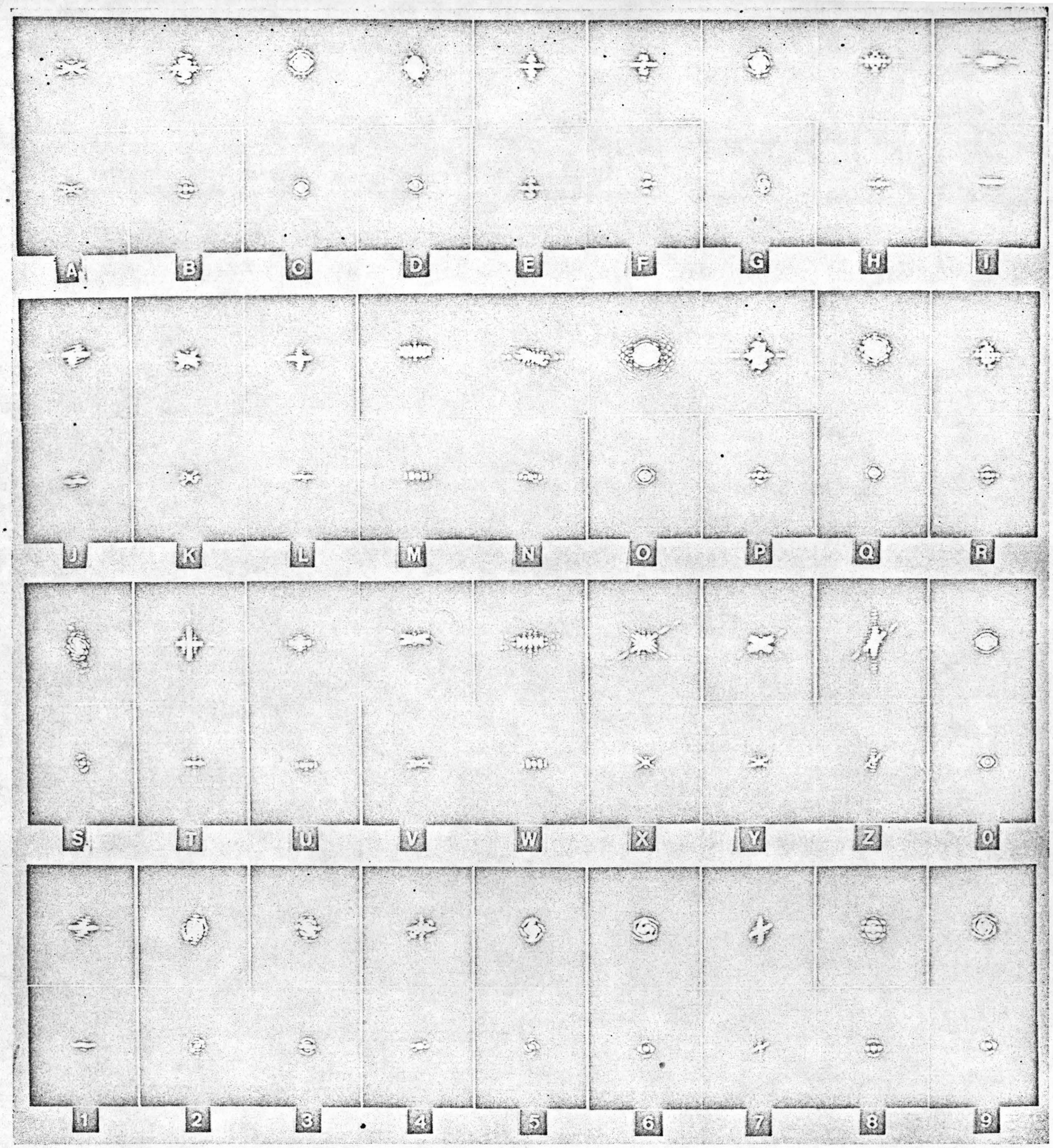


Fig. 8. Fraunhofer patterns of entire alphanumeric character font, recorded at two different exposures to show both the low and high spatial frequency regions.

thin metal sheet, forming a miniature stencil mask as shown in Fig. 6. Since the transparent parts of the characters were completely open, there was no chance for phase effects to be introduced, and the alignment of the system was independent of the presence of the character. By scanning the output diffraction spectra and looking for any radial asymetry, very small alignment errors can easily be detected.

The appearance of several diffraction spectra at the output plane are shown in Fig. 7. The spectra are shown, with and without the spot, for a 1-mm circular aperture and the characters *A* and *B*. There

U:U	1.00	O:O	1.00
U:O	.86	O:Q	.96
U:W	.75	O:D	.91
U:Z	.43	O:C	.85

(a) (b)

Fig. 9. Numeric values for correlation coefficients: (a) typical combinations; (b) difficult combinations.

was no filter used. The entire set of alphanumeric characters is shown in Fig. 8, with the spectra recorded at two different exposures to show both the low and high spatial frequency parts of the spectra.

Figure 9 shows correlation coefficients for two different sets of four characters each, chosen to represent four typical and four difficult combinations of characters. These numerical data, taken with the spot in position, indicate the discrimination achieved in a laboratory experiment.*

The sensitivity of this experimental system to shifting of the input pattern was also investigated and found to be negligible over the field measured (ten character widths by ten character heights). The sensitivity to rotation[15] is somewhat less than is the case for direct character correlation. Preliminary experiments and computations indicate that rotations should not exceed about 5°. These experiments have shown also that parallel processing by theta-modulating the known characters does not present any unexpected problems. The ratio of character line-width to grating spacing should be about 3 or greater to avoid overlapping diffraction patterns.

IX. Conclusions

A new method of optical character recognition has been described, involving three principal stages:

(a) Preprocessing the unknown character $u_n(x_0)$ to obtain the incoherent shift invariant signal $I_n(x)$.

(b) Realization of a set of incoherent matched filters $T_m(\nu) = \tilde{I}_m^*/S_{mm}$ such that the correlation coefficients measured will satisfy the recognition inequality $S_{nn} > S_{nm}$ $(m \neq n)$, permitting recognition.

(c) Application of the set of incoherent matched filters, and measuring the intensity at $x = 0$ at the output plane to obtain the set of S_{nm}.

Also described is a simple method which increases discrimination when applied to the matched filter input, and an elaboration to permit parallel processing. Some experiments are reported which demonstrate the feasibility of this incoherent matched filtering method.

Compared to the four other most well-known methods of optical character recognition, those of direct character correlation, correlation of diffraction spectra, coherent matched filtering, and shadow correlation, this new method is more complicated to describe mathematically. However, it is the only one of these methods which combines both of the practical advantages: shift invariance and only opaque-transparent masks. These advantages are gained by devising a matched filter system which can be applied to *intensity* rather than *amplitude* inputs, and which is correct within scalar diffraction theory.

We acknowledge with pleasure the patient and skilful experimental contributions of Harald Werlich, and stimulating discussions with Ronald Kay and Dieter Paris. We also would like to thank A. Vander Lugt and A. Kozma for clarifying discussions.

References

1. L. P. Horwitz and G. L. Shelton Jr., Proc. Inst. Radio Engrs. **49**, 175 (1961).
2. E. Baumann, J. Soc. Motion Picture Television Engrs. **60**, 344 (1953).
3. A. Maréchal, in *Communication and Information Theory in Optics* (G. E. Co., Syracuse, N. Y., 1960); E. O'Neill, IRE Trans. **IT-2**, 56 (1956); J. Tsujiuchi, *Progress in Optics* (Wiley, New York, 1963), Vol. II, L. Cutrona *et al.*, Inst. Radio Engrs. **IT-6**, 391 (1960).
4. A. Lohmann, Optica Acta **5**, 293 (1958); **6**, 319 (1959); in *Communication and Information Theory in Optics* (G. E. Co., Syracuse, N. Y., 1960).
5. E. Leith and J. Upatnieks, J. Opt. Soc. Am. **53**, 1377 (1963).
6. A. Vander Lugt, IEEE Trans. **IT-10**, 140 (1964).
7. D. H. Kelly, J. Opt. Soc. Am. **51**, 1095 (1961).
8. E. A. Trabka and P. G. Roetling, J. Opt. Soc. Am. **54**, 1242 (1964).
9. A. Lohmann and H. Wegener, Z. Physik **143**, 431 (1955); IBM Tech. Rept., available from the authors on request.
10. G. Toraldo di Francia, *Electromagnetic Waves* (Interscience, New York, 1953).
11. P. M. Duffieux, *L'Intégral de Fourier et ses Applications à l'Optique* (Chez l' Auteur, Université de Besançon, Besançon, 1946); H. H. Hopkins, Proc. Roy. Soc. (London) **A217**, 408 (1953).
12. J. Armitage and A. Lohmann, Appl. Opt. **4**, 399 (1965).
13. E. Leith, Phot. Sci. Eng. **6**, 75 (1962).
14. D. Delwiche, J. Clifford, and W. Weller, J. Soc. Motion Picture Television Engrs., **67**, 678 (1958).
15. D. Paris, IBM Tech. Rept., available from the author on request.

* *Note added in proof.* A later more careful theoretical analysis of the requirements on the diffuser shows that the experimental data reported here (Fig. 9) are only of order-of-magnitude significance. The lateral and longitudinal grain size distribution of the diffuser as used in this experiment must permit $I_n(x)$ to be equivalent to a self-luminous object in the time average. How to achieve this, especially with the very long coherence lengths of laser light, is not a simple problem.

Reprinted from APPLIED OPTICS, Vol. 4, page 445, April 1965

Absolute Contrast Enhancement

J. D. Armitage, A. W. Lohmann, and R. B. Herrick

The basic idea of photographic masking is discussed, and two new methods for producing such masks are described. The conditions necessary to justify application of modulation transfer theory to the process are specified, and the theory is applied to demonstrate that it is possible to obtain a modulation transfer function greater than unity for a band of spatial frequencies. Experimental verification is described.

Introduction

The unsharp masking technique may be considered in two ways: density compression to accommodate materials of short exposure latitude (such as in photographic printing), and as a form of spatial filtering.[1,2] We will consider one aspect of the latter. Spatial filtering is often used to accomplish edge-sharpening (contour enhancement), or, in less photographic terms, to increase the amplitude of the higher (spatial) frequency Fourier spectra in the output image. As unsharp masking is usually performed, this increase in the higher frequency modulation is an increase only relative to the lower frequency modulation. It is not an absolute increase. This relative increase is obtained by decreasing the lower frequency modulation relative to that of the higher frequencies, and, in actuality, even the higher frequency modulation may be somewhat decreased. We will show here that it is possible to achieve an *absolute* increase in modulation for a band of frequencies. In modulation transfer theory, this means a modulation transfer function (MTF) greater than unity. Since modulation transfer theory is usually applied to linear systems with non-negative signals, and since under the condition of non-negative signals an MTF greater than unity is impossible, an absolute increase in modulation such as we will discuss here has usually been considered impossible.[3,4] We will show analytically and experimentally that the masking process may justifiably be approximated as a linear process under suitable conditions, and that under these conditions of quasi-linearity a MTF greater than unity is possible.

Usual and Modified Masking Techniques

Assume that the input to the system is a photographic transparency with a linear grating pattern varying sinusoidally in transmittance T_i along the x direction (Fig. 1). There are two spatial frequencies, a lower frequency R_1 and a higher frequency R_2. By some special method of photographic printing, from this input transparency we produce the mask transparency which is the same size as the input and with a transmittance distribution $T_m(x)$ resembling, but not equaling, that of the input. The input and mask transparencies are placed in register, and printed to form the output of transmittance distribution $T_o(x)$, equal point-by-point to the product $T_i(x) \cdot T_m(x)$.

In the usual unsharp masking process, the mask is formed by slightly defocusing the enlarger lens or by contact printing with a thin spacer between the emulsions. This results in a *blurring* which is more severe for higher frequencies, affecting the transmittance of the mask and output as shown in the left-hand part of Fig. 1. In this case the mask is simply a nega-

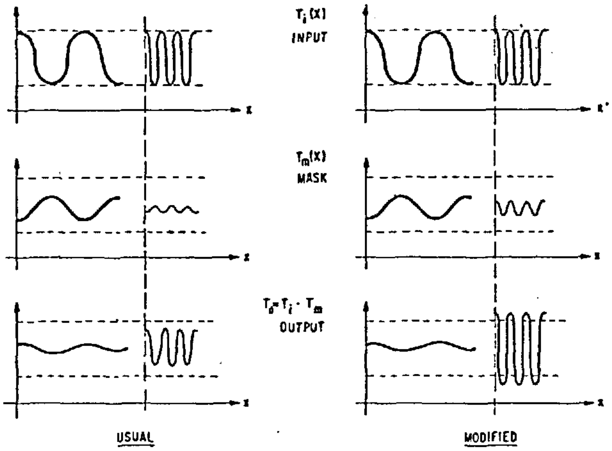


Fig. 1. Schematic representation of *usual* and *modified* masking techniques, using patterns sinusoidal in transmittance.

J. D. Armitage and A. W. Lohmann are with IBM Corporation, General Products Division Development Laboratory, San Jose, California. R. B. Herrick is with Utah State University, Logan, Utah.

Received 14 June 1964.

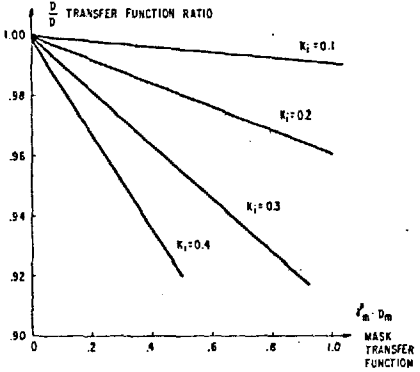


Fig. 2. Plot of ratio of approximate to exact transfer functions vs mask transfer function, for several values of input contrast. Variable M in the text is written in this figure as variable K.

tive copy of the input, with higher frequencies reproduced at lower modulation.

If it were possible, in a modified method for producing the mask, to achieve a second reversal for a specified range of frequencies, as is shown on the right-hand side of Fig. 1, the masking procedure would increase rather than decrease the modulation for these frequencies.

Application of Modulation Transfer Theory

Linearity and Nonnegative Signals

To be able to describe the masking process in terms of modulation transfer theory, the operations involved must be at least approximately linear. Let the input transmittance function be $T_i = a_i(1 + K_i \cos 2\pi Rx)$ and that of the mask be $T_m = a_m(1 + K_m \cos 2\pi Rx)$, where the modulation $M = (T_{\max} - T_{\min})/(T_{\max} + T_{\min})$. The output transmittance function T_o will be the product of $T_i \cdot T_m$ since the transmittances multiply

$$T_o = T_i(x) \cdot T_m(x) =$$

$$a_i a_m[1 + (M_i + M_m)\cos 2\pi Rx + M_i M_m \cos^2 2\pi Rx].$$

The last term represents the harmonic generation introduced by the nonlinearity of the multiplicative process. To approximate the process as being linear, the coefficient of this term relative to that of the strictly linear preceding term should be small:

$$(M_i M_m)/(M_i + M_m) \ll 1, \text{ or } (1/M_i) + (1/M_m) \gg 1.$$

Restricting the modulation in either the input or mask transparencies will achieve the desired minimization of the nonlinear coefficient. A numerical example will indicate the range of modulation permissible. Let $M_i = 0.3$ and $M_m = 0.2$;

$$T_o(x) = a_i a_m[1 + 0.5 \cos 2\pi Rx + 0.06 \cos^2 2\pi Rx].$$

So, $(M_i M_m)/(M_i + M_m) = 0.12 \ll 1$. If modulation is restricted to these ranges, the process may be regarded as linear to about 10%, and, within this linearity, modulation transfer theory may be applied.

In a strictly linear system, if the input modulation $M_i = 1.0$ and the modulation transfer function $D(R) > 1.0$, then $M_o > 1.0$, which is not reasonable since it implies negative intensity. To have $D(R) > 1.0$, M_i must be considerably less than 1.0. Fortunately, this is already assured from harmonic distortion requirements.

MTF of Mask Production $D_m(R)$ and MTF of Entire System $D(R)$

How is the system transfer function $D(R)$ related to the mask production transfer function $D_m(R)$? Is it possible to specify the characteristics of $D_m(R)$ necessary to produce a desired $D(R)$?

As before, assume an input grating sinusoidal in transmittance, with maximum transmittance $T_{i-\max}$ and minimum transmittance $T_{i-\min}$. Let $(T_{i-\min})/(T_{i-\cancel{\min}\,\max}) = \epsilon_i$.

Then,

$$M_i = \frac{T_{i-\max} - T_{i-\min}}{T_{i-\max} + T_{i-\min}} = \frac{1 - (T_{i-\min}/T_{i-\max})}{1 + (T_{i-\min}/T_{i-\max})} = \frac{1 - \epsilon_i}{1 + \epsilon_i}.$$

The mask may be produced using a photographic material having a gamma γ_m independent of frequency and a MTF equal to 1.0. The mask will be sinusoidal (or quasi-sinusoidal because of possible harmonic distortion) resembling the input, and it will have its corresponding $T_{m-\max}$ and $T_{m-\min}$, which determine the mask modulation $M_m = (1 - \epsilon_m)/(1 + \epsilon_m)$.

The mask production transfer function is defined by

$$D_m(R) = M_m(R)/[M_i(R) \cdot \gamma_m].$$

Now combine the mask and input in register so the $T_{i-\max}$ falls on $T_{m-\min}$ and vice versa.

$$T_{o-\max} = T_{i-\max} \cdot T_{m-\min} \text{ and } T_{o-\min} = T_{i-\min} \cdot T_{m-\max}.$$

$$\epsilon_o = T_{o-\min}/T_{o-\max} = \epsilon_i/\epsilon_m.\, M_o = (1 - \epsilon_o)/(1 + \epsilon_o).$$

Although the output transmittance distribution may no longer be sinusoidal, $T_{o-\max}$ and $T_{o-\min}$ may be measured at the 0 and π points of the input sinusoid corresponding to its maximum and minimum. The transfer function for the entire masking procedure

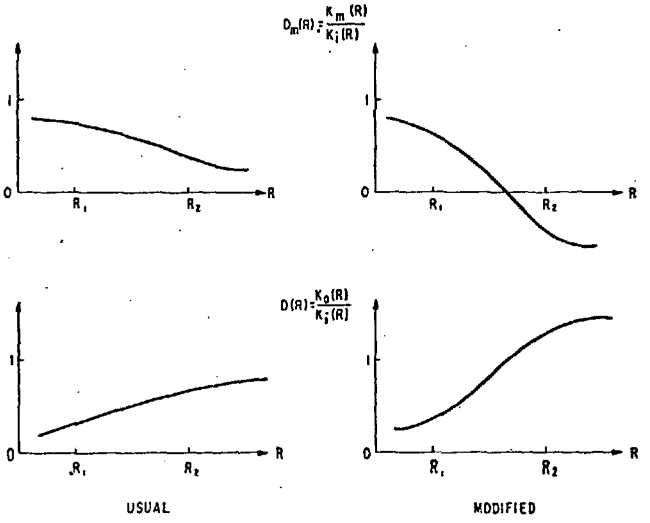


Fig. 3. Mask and system transfer functions for the *usual* and *modified* methods. Variable M in the text is written in this figure as variable K.

METHOD	INPUT	MASK
DEFOCUS	•	●
LATERAL SHIFT	•	• •
RING SMEAR	•	○

Fig. 4. Transformation of a point in the input into the appropriate pattern in the mask, for three different masking methods.

(system transfer function) $D(R) = M_o(R)/M_t(R)$. (For convenience, omit the argument R in the following development.)

$$M_i \cdot D = M_o = \frac{1 - \epsilon_o}{1 + \epsilon_o} = \frac{1 - (\epsilon_i/\epsilon_m)}{1 + (\epsilon_i/\epsilon_m)} = \frac{\epsilon_m - \epsilon_i}{\epsilon_m + \epsilon_i}$$

$$= \frac{1 - \epsilon_i}{1 + \epsilon_i} - \left[\frac{1 - \epsilon_i}{1 + \epsilon_i} - \frac{\epsilon_m - \epsilon_i}{\epsilon_m + \epsilon_i}\right].$$

Examine the expression in the brackets[].

$$[] = M_i - \frac{(1 - M_m)/(1 + M_m) - (1 - M_i)/(1 + M_i)}{(1 - M_m)/(1 + M_m) + (1 - M_i)/(1 + M_i)}$$

$$= M_i - \frac{(1 - M_m)(1 + M_i) - (1 - M_i)(1 + M_m)}{(1 - M_m)(1 + M_i) + (1 - M_i)(1 + M_m)}$$

$$= M_i - \frac{(1 - \gamma_m D_m M_i)(1 + M_i) - (1 - M_i)(1 + \gamma_m D_m M_i)}{(1 - \gamma_m D_m M_i)(1 + M_i) + (1 - M_i)(1 + \gamma_m D_m M_i)}$$

$$= M_i - \frac{2M_i(1 - \gamma_m D_m)}{2(1 - M_i^2 \gamma_m D_m)}$$

$$= \frac{M_i(1 - M_i^2\gamma_m D_m) - M_i(1 - \gamma_m D_m)}{1 - M_i^2\gamma_m D_m} = \frac{M_i \gamma_m D_m (1 - M_i^2)}{1 - M_i^2 \gamma_m D_m}.$$

So,

$$M_i \cdot D = M_i - []; \ D = 1 - \frac{[]}{M_i}$$

$$= 1 - \gamma_m D_m \left(\frac{1 - M_i^2}{1 - M_i^2 \gamma_m D_m}\right). \quad (1)$$

It would obviously be convenient to approximate $D(R)$ by $\tilde{D}(R) = 1 - \gamma_m \cdot D_m(R)$. The constraint that $M_t \ll 1$, required for approximate linearity, justifies this approximation by making the term $[(1 - M_t^2)/(1 - M_t^2\gamma_m D_m)]$ go to unity. Figure 2 shows the error in this approximation, as a function of $\gamma_m D_m$ for four appropriate values of M_t and foı $\gamma_m = 1.0$. The approximation is certainly reasonable.

With the approximation $D(R) = 1 - \gamma_m \cdot D_m(R)$, it is apparent that to obtain a $D(R)$ greater than unity, $\gamma_m \cdot D_m(R)$ should be negative. Such negative transfer functions are known, and the effect is often termed "spurious resolution".[3]

In Fig. 1, the masking process is described schematically, showing the spatial transmittance functions of the input, mask, and output. Figure 3 is similar, but in the frequency domain rather than in the space domain, showing the transfer functions for the usual and modified masking techniques.

Mask Production

Methods

Consider three possible methods for producing the mask from the input:

1. defocusing (as in the usual unsharp masking),
2. lateral shift with double exposure,
3. ring smearing.

Figure 4 shows, for each of these three methods, the transformation of an ideally small point in the input to the appropriate shape in the mask. The first method, defocusing, should be self-explanatory. In the second, lateral shift with double exposure, the mask is a double-exposed copy of the input with a shift in the x direction between the exposures. In the third, ring smearing, the mask (or input) is moved in a circular path during the exposure so that each point in the input is smeared into a circle (not a disk) on the mask.

Transfer Functions

To calculate the transfer function $D_m(R)$ for the mask production process, consider the intensity distribution $\Phi(x,y)$ produced from a point object in the input plane. The transfer function is

$$D_m(R_x,R_y) = \frac{\iint \Phi(x,y) \exp[-2\pi i(xR_x + yR_y)]dxdy}{\iint \Phi(x,y)dxdy}, \quad (2)$$

where the denominator normalizes the transfer function to unity at zero frequency.

Defocused Transfer Function

Figure 5(a) shows schematically the physical setup for the production of the mask by the usual defocusing method. With a single infinitesimal point input, assume a uniform intensity distribution in the defocused image in the mask plane. If diffraction effects can be neglected, the derivation of the transfer function is straightforward.[5] If the radius ρ of the defocused disk is sufficiently greater than the radius of the Airy disk ρ_A in the Gaussian focal plane, the Airy disk can be assumed infinitesimal also and a strictly geometrical treatment will be valid. Assume that a factor of 3 is sufficient. $\rho_A = (1.22\,\lambda b)/(2h)$, and $\rho = hz/b$. Therefore, for $\rho = 3\rho_A$, we have $(hz/b) \geq (3.66\,\lambda b)/(2h)$, or $z \geq (3.66\,\lambda b^2/2h^2)$. Substituting typical experimental values, $\lambda = 5 \times 10^{-4}$ mm, $b/h = 16$, we have $z \geq 0.234$ mm. Therefore, if we defocus by at least this amount, we may approximate the process as a geometrical one.

The transfer function $D(R,z)$ is the Fourier transform of the intensity distribution function $\Phi(x,y)$. Since this function is a flat-topped cylinder in the out-of-focus plane, this calculation is the same as that for the Airy disk diffraction pattern, resulting in $D(R,z) = 2J_1(2\pi\rho R)/(2\pi\rho R)$. Figure 5(b) is a plot of $D(R,z)$ vs $2\pi\rho R$. The transfer function is negative as desired

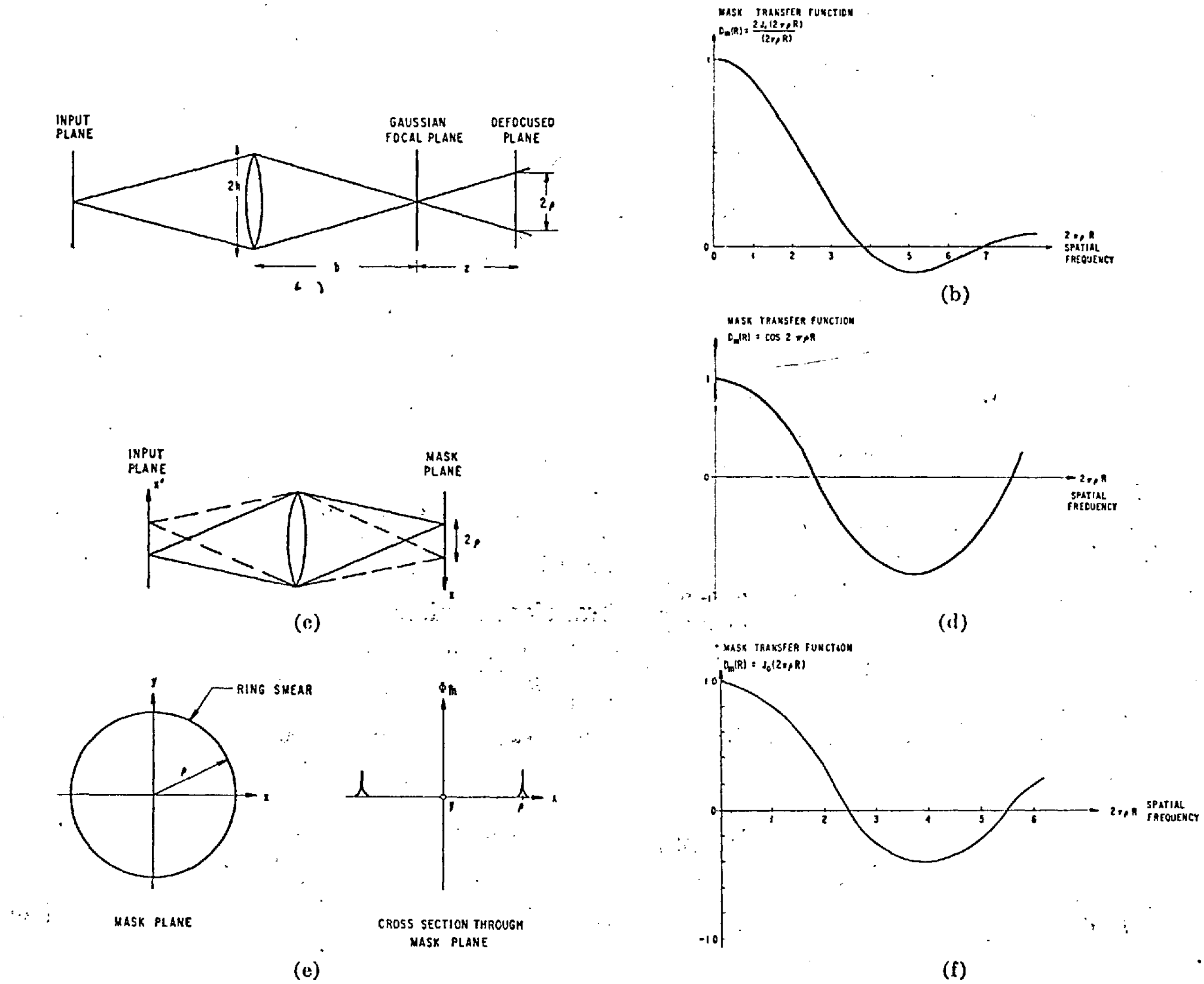


Fig. 5. Schematic diagram of mask production by defocusing method. (a) Mask production by defocusing method. (b) Mask transfer function for defocusing case. (c) Mask production for lateral shift with double exposure method. (d) Mask transfer function for lateral shift case. (e) Appearance of the mask plane for a point in the input plane. (f) Mask transfer function for ring smear case.

over a certain frequency band; however, it goes negative only to about − 0.1, which is for practical purposes not strongly negative.

Transfer Function of Lateral Shift with Double Exposure

Figure 5(c) shows schematically the experimental setup for producing the mask by this method. The input is imaged onto the mask plane, an exposure is made, the image is shifted relative to the mask by 2ρ in the x direction, and a second equal exposure is made. This method is obviously appropriate for objects having structure only along the x direction.

Let the intensity after the input transparency be

$$\Phi_i(x) = {}^1/_2[1 + M_i \cos(2\pi Rx)].$$

After the two exposures, the exposure at the mask will be

$$\Phi_m(x) = {}^1/_4\{1 + M_i \cos[2\pi R(x + \rho)]\} + {}^1/_4\{1 + M_i \cos[2\pi R(x - \rho)]\}$$
$$= {}^1/_2[1 + M_i \cos(2\pi Rx) \cos(2\pi R\rho)].$$

By definition, $D_m(R)$ = mask contrast/input contrast: $D_m(R) = \cos 2\pi\rho R$, as shown in Fig. 5(d). Note that the transfer function goes strongly negative, to a possible −1.0.

Transfer Function for Ring Smearing

Let the input intensity function be a point, so that $\Phi_i(x,y) = \delta(0,0) = \Phi_i(r,\phi)$. The mask function will be $\Phi_m(r,\phi) = \delta(r - \rho)$, independent of ϕ as shown in Fig. 5(e). As with the case of the defocused transfer function, it follows from Eq. (2) that the transfer function

$$D_m(R,\rho) = \frac{\int_0^{2\pi}\int_0^{\rho} \Phi_m(r,\phi) \exp[-2\pi iRr \cos(\phi - \psi)] r dr d\phi}{\int_0^{2\pi}\int_0^{\rho} \Phi_m(r,\phi) r dr d\phi},$$

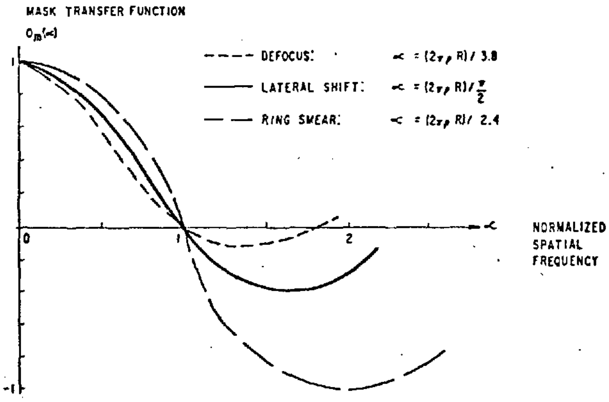


Fig. 6. The three mask transfer functions plotted on normalized frequency coordinates.

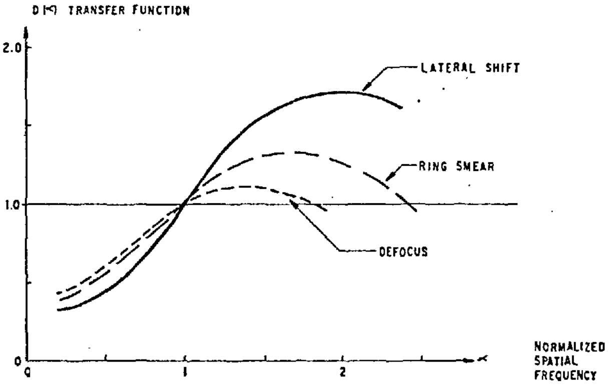


Fig. 7. The three system transfer functions plotted on normalized frequency coordinates.

$$D_m(R,\rho) = \frac{\int_0^{2\pi}\int_0^{\rho} \delta(r-\rho)\exp[-2\pi iR\cos(\phi-\psi)]rdrd\phi}{\int_0^{2\pi}\int_0^{\rho}\delta(r-\rho)rdrd\phi},$$

$$D_m(R,\rho) = \frac{\int_0^{2\pi}\exp[-2\pi iR\rho\cos(\phi-\psi)]\rho d\phi}{\int_0^{2\pi}\rho d\phi} = \frac{\rho 2\pi J_o(2\pi\rho R)}{\rho 2\pi}.$$

We have $D_m(R,\rho) = J_o(2\pi\rho R)$, which is plotted in Fig. 5(f). Note that although this transfer function is not as strongly negative as that for the lateral shift method, it is considerably more strongly negative than for the defocused case.

System Transfer Function

The three mask transfer functions D_m are plotted in Fig. 6 on a normalized frequency axis so that they cross the frequency axis at a common point. The ratio M_m/M_i in the case of the lateral shift case has been set equal to unity. The dummy frequency variable is α with the following normalizations:

Defocused: $\alpha = (2\pi\rho R)/3.8$,
Lateral shift: $\alpha = (2\pi\rho R)/0.5\pi$,
Ring smear: $\alpha = (2\pi\rho R)/2.4$.

With these mask transfer functions, the approximations $D(R) = 1 - \gamma_m D_m(R)$ and $\gamma_m = 0.7$, the system transfer functions $D(R)$ for the three masking methods are plotted in Fig. 7. Over the appropriate frequency bands the transfer function $D(R)$ is significantly greater than unity, especially in the cases of lateral shift and ring smear.

Experimental

In order to verify experimentally the above theoretical points, we performed photographic masking experiments using the lateral shift and ring smear methods for mask production. The defocusing method was not

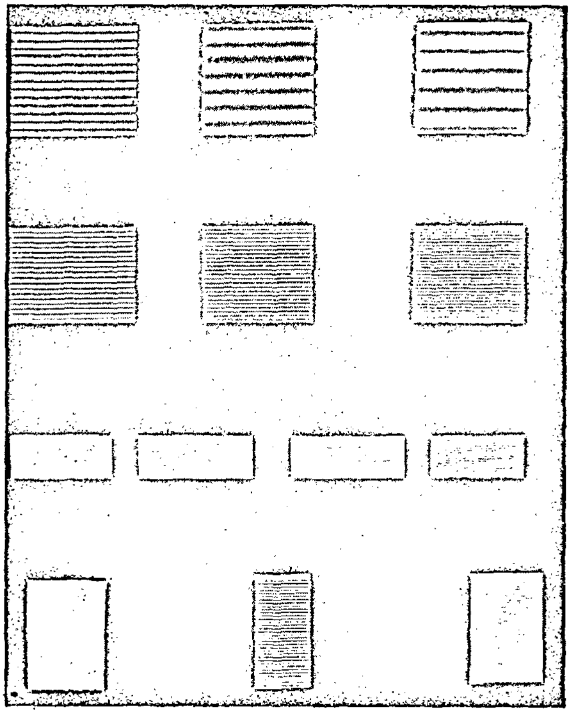


Fig. 8. The input sine wave pattern.

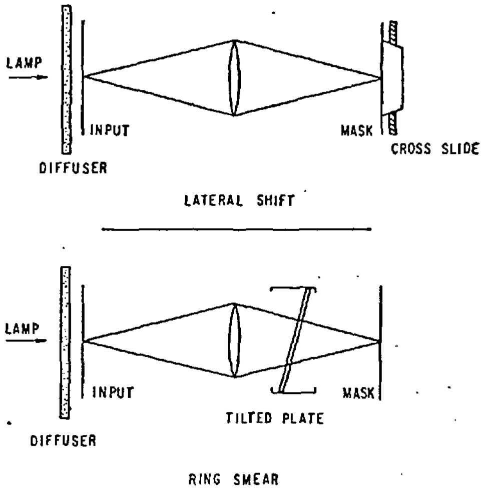


Fig. 9. Diagram of experimental setups for lateral shift and ring smear mask production.

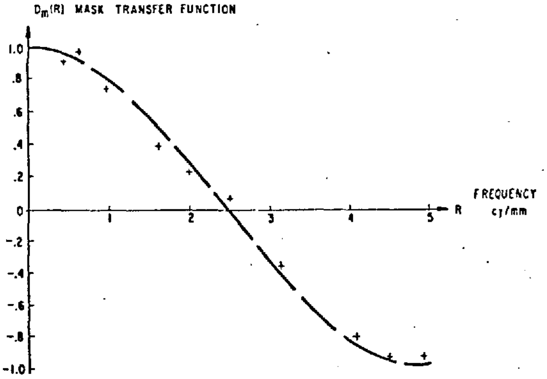


Fig. 10. Mask transfer function for lateral shift case, showing experimental points and theoretical curve.

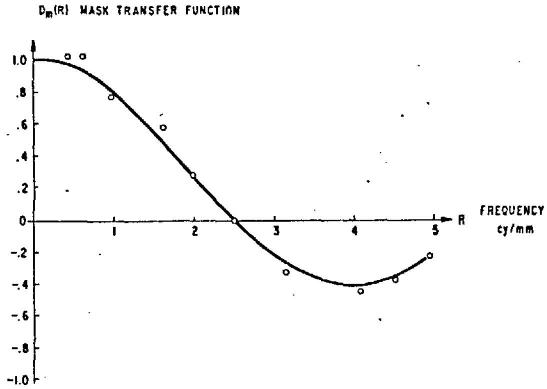


Fig. 11. Mask transfer function for ring smear case, showing experimental points and theoretical curve.

used because of its low (≈13% maximum) contrast enhancement. As input, we used a group of ten linear sinusoidal transmittance targets (Fig. 8) of frequencies 0.411, 0.494, 0.971, 1.60, 1.99, 2.50, 3.11, 4.07, 4.51, and 4.95 cycles/mm, with an average modulation 0.7 and average density 0.8. These rather low spatial frequencies were chosen to minimize mechanical experimental difficulty and to enable us to assume the MTF of lenses and photographic materials used equal to unity. (In reality, considerable care must be taken to avoid the adjacency effect,[6] which can make the latter assumption a dangerous one.) Square-wave line patterns were placed at the edges of the input target to facilitate determination of ρ to give the desired zero-crossing of the mask transfer function and the final alignment of the transparencies in registration.

The input transparency was diffusely illuminated with a Graflarger system, and projected with a high quality enlarging objective onto Panatonic-X film developed to unit gamma. The resulting densities were controlled to fall on the *linear* part of the H & D curve of the film. Figure 9 shows schematically the setup used. In the lateral shift case, the photographic film was translated laterally with a precision cross-slide with two exposures being made at the proper shift distance. In the ring smear case, a thin glass plate tilted relative to the optical axis and placed in the convergent beam was rotated rapidly several times uring the exposure to produce the desired circular image motion. To compensate for uncontrolled experimental variables, additional transparencies were produced under the same conditions but without the lateral shift or glass plate rotation. With a mask modulation $M_m(R)$ and control modulation $M_c(R)$ the mask transfer function $D_m(R)$ is $D_m(R) = M_m(R)/M_c(R)$; the control transparencies thus serve to normalize the mask production process. The transparencies were scanned on a Joyce double-beam microdensitometer to determine the modulations and transfer functions, which are shown in Figs. 10 and 11. For both cases, the parameter ρ was chosen to have the mask transfer function $D_m(R) = 0$ at 2.5 cycles/mm.

To avoid obscuring the effects of interest in the experiment with other effects not germane, such as adjacency effect, photographic nonlinearities, and lens imperfections, we decided to separate the masking procedure into two parts: the production of the mask and the production of the output once the mask has been obtained. This makes the experiment less close to real practice, but permits closer control for demonstrating the validity of the theoretical treatment. For the second part of the process, the control transparency was used as input. It and the mask were contact-printed

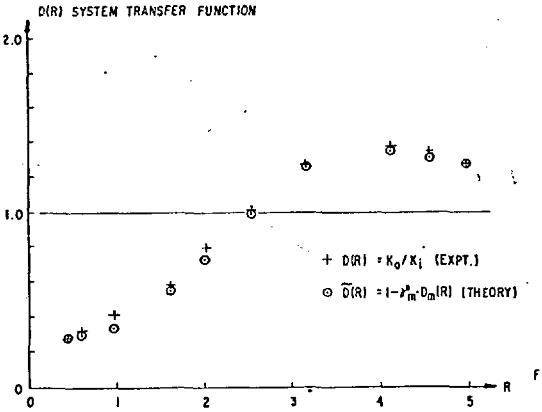


Fig. 12. System transfer function for lateral shift case, showing experimental and theoretical points.

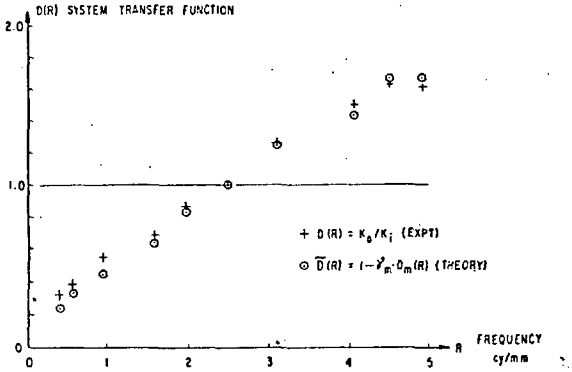


Fig. 13. System transfer function for ring smear case, showing experimental and theoretical points.

onto Plux-X Pan film developed to a γ of approximately 0.5 to reduce the modulation to an acceptable level. The mask was then contact-printed again at unit gamma to produce the correct polarity. These final transparencies were considered as the input and mask for the system, and their superposition in register considered the output. If the mask modulation is $M_m(R)$, the input modulation $M_i(R)$, and the output modulation $M_o(R)$, then the transfer functions will be $D(R) = M_o(R)/M_i(R)$ and $\tilde{D}(R) = 1 - \gamma_m \cdot D_m(R)$. Because of the normalization by the control, $\gamma_m = 1$. Figs. 12 and 13 show the experimental transfer functions compared with those predicted by the approximative theory.

Conclusions

Although in the linear systems approach in optics one would not normally speak of a modulation transfer function greater than unity, with reasonable approximations it is realistic to consider the possibility of applying linear theory to nonlinear systems, and to consider the possible usefulness of doing this. The photographic masking methods described in this paper represent one simple application of such thinking, and it is hoped that others may be stimulated to consider similar somewhat unusual applications of transfer theory to optical problems.

The authors would like to thank H. Werlich for his excellent performance of much of the experimental work described, and H. Zweig for helpful critical discussion. M. Rabedeau supplied the sine targets, which simplified our experimental problem considerably.

References

1. A. Lohmann, Opt. Acta **6**, 319 (1959).
2. A. Lohmann, *Communications and Information Theory Aspects of Modern Optics* (General Electric Co., Syracuse, N. Y., 1962), p. 51.
3. H. Frieser, Appl. Opt. **3**, 15 (1964).
4. W. Spitzberg and F. A. Sunder-Plassmann, Optik **20**, 440 (1963).
5. P. Lindberg, Opt. Acta **1**, 80 (1954).
6. L. Hendeberg, Arkiv Fysik **16**, 457 (1960).

Reprinted from APPLIED OPTICS, Vol. 4, page 399, April 1965

Theta Modulation in Optics

J. D. Armitage and A. W. Lohmann

The experiments reported in this paper are similar to the famous Abbe experiments. However, they were done for quite different reasons, namely, to perform certain information processing operations by optical means. Our technique, called theta modulation, allows production of a color image from a black and white film, on which the color object is recorded in encoded form. Furthermore, nonlinear characteristics (H & D curves) of any shape can be realized. A special application of theta modulation, called multiplex storage, will be described. By this technique, more than one image can be recorded in the same area on a piece of film. Subsequently, the individual images can be recovered with a minimum of *crosstalk*.

I. The Principle of Theta Modulation

The meaning of modulation is well known in electronics. One has, for example, a sinusoidal carrier and a signal, and both are brought together such that the amplitude (or frequency or phase) of the carrier varies proportionally to the signal. In this modulated form, the signal may be easier to transmit. Or, several signals riding on different carriers might be transmitted simultaneously over one single cable. This technique, called "multiplexing", will be one of the applications for our optical modulation scheme. Whereas in electronics (and also in laser communications systems) the signal is a temporal function, let us say voltage as function of time t, we here mean by signal a spatial function, an intensity distribution $I(x,y)$ or $I(r,\phi)$. By "carrier" we mean a *spatial monofrequency*, or, in more common terms, an amplitude grating. It is obvious that one could modulate such a carrier essentially the same way as in electronics by varying its amplitude, frequency, or phase proportional to the signal. Thermoplastic recording[1] falls into this category. But there is another parameter suitable for modulation, and this parameter is unique in our optical situation; this is the angular orientation of the grating, here defined as the azimuth angle theta. This parameter has no counterpart in electronics, where one has only one independent variable, the time t. Figure 1 shows what we mean by theta modulation. At the left side of the figure, one sees a signal, $I_0(x,y)$ at the right side of the modulated signal $I_M(x,y)$. Modulation here means the assignment of various theta angles of the carrier to corresponding intensity levels.

To implement the modulation in the experiments here described, we simply used commercial grid paper (Ronchi ruling), scissors, and glue. A more useful way to perform the modulation will be described in the Appendix.

Now, let us describe the demodulation process, first for the simplest case and in the following sections, with some of the features mentioned in the abstract. The basic setup is a coherent image-forming system (Fig. 2). If the lines of the grating in M are oriented vertically, the diffraction patterns are situated on a horizontal line in the Fraunhofer plane F, as indicated in the upper part of Fig. 3. If the grating is rotated by an angle theta, the diffraction orders will rotate by the same amount around the center of F. If the object consists of two gratings, side by side with different azimuthal orientations, the diffraction pattern will look as shown at the bottom of Fig. 3. Now, assume an even more general object as in Fig. 1(b) or Fig. 4. In the diffraction plane, the several diffraction patterns appear in discrete angular positions, according to the angular orientation of the grating elements. Hence, if one wants the roof of the house to appear dark in the image plane B (see Fig. 2), one has to block out the diffraction orders which correspond to the roof. A light sky can be achieved by letting pass both diffraction orders of the sky. The wall will be grey if only one of the two *wall diffraction spots* passes through the *demodulation mask* as indicated at the right-hand side of Fig. 4. The zero diffraction order at the center of F is blocked out at all times, since it is not dependent on the grating angle theta.

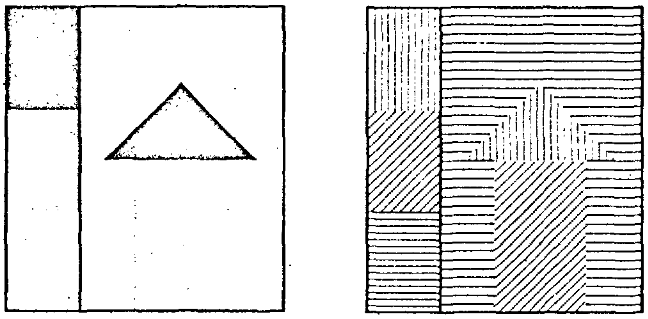
Results of this type of demodulation experiment are presented in Fig. 5. Two different images could be achieved, depending upon which demodulation mask was used.

These qualitative arguments readily suggest to anyone familiar with Fourier optics the manner in which

The authors are with the IBM Corporation, San Jose, California.

Received 17 November 1964.

This paper was presented at the OSA meeting in Chicago, 23–25 October 1963.



(a). (b)

Fig. 1. Principle of theta modulation. (a) Object with grey ladder; (b) same object in theta-modulated form.

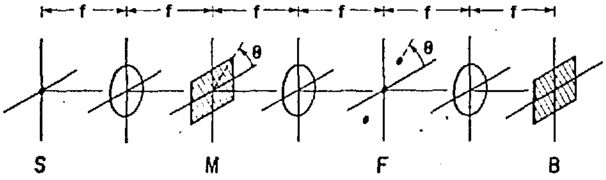


Fig. 2. Optical arrangement for theta demodulation. S = source; M = plane for modulated object; F = Fraunhofer plane, the place for the demodulation mask; B = image plane, where the demodulated object appears.

the theory of our experiments would be developed. Some details which would benefit from quantitative treatment have been investigated theoretically and prepared for publication in *Optik* by Morgenstern.

II. Production of Color Images

Now we will describe three ways to produce a color image from a theta-modulated signal which consists of black and white amplitude gratings. First, one can split up each object element into three portions, as shown for four object elements in Fig. 6. The mask consists of opaque and transparent portions, superimposed by blue, red, and green filters (B, R, and G, in Fig. 6). For example, the lower third of each object element always is responsible for the blue component. There the grating angle theta varies between zero and $\pi/3$. In this simple mask (Fig. 6), only three amplitude levels for each color component, say zero, 0.5, and 1, can be generated. Continuous values are possible, of course, if the transmission of the mask varies continuously as a function of the angle θ.

In a second version of these color image experiments, one also uses three elementary gratings for the three color components in each object element, however, not side by side but superimposed. Qualitatively, this works satisfactorily when using similar demodulation masks as before (i.e., Fig. 6). However, for a quantitatively true color reproduction, the mask must eliminate the moiré diffraction orders which are due to multiplicative interaction of the superimposed gratings.

Our best results were achieved with a third version of these experiments. Here we make use of the fact that different wavelengths appear in the diffraction plane at different radii. If, for example, we want to generate a color image of our house object (Figs. 1, 4, 5), we insert a black and white mask (Fig. 7) into the Fraunhofer plane, such that only the *blue* portion of the *sky diffraction spots* can go through, and so on. A somewhat more general color mask in which the transmitted wavelength is proportional to the angle theta of the grating would consist of a spiral slit in an opaque mask, given a polychromatic source.

III. Nonlinear Demodulation

Let us assume a linear modulation of the input. That is, the original object, $I_0(x,y)$, is converted into a modulated signal $I_M(x,y)$ with a local grating angle θ, proportional to the intensity distribution in the object:

$$\theta(x,y) = KI_0(x,y), \qquad K = \pi/\max(I_0).$$

For the demodulation mask, situated in the Fraunhofer plane F of Fig. 2, let us assume a transmission $T(\theta)$, which is not necessarily a linear function of the azimuth angle theta. Hence, the image intensity, $I_B(x,y)$, may be related in a nonlinear or even nonmonotonic manner to the object intensity $I_0(x,y)$. The following scheme summarizes the described procedures:

$$\begin{array}{lr} \text{modulation:} & I_0 \rightarrow \theta = KI_0, \\ \text{demodulation:} & \theta \rightarrow T(\theta) = I_B, \\ \text{total:} & I_0(x,y) \rightarrow I_B(x,y) = T[KI_0(x,y)]. \end{array}$$

In other words, the angular variation $T(\theta)$ of the demodulation mask may act as the nonlinear characteristic, much the same way as the H & D curve of a photographic material influences an optical signal. However, now we have the ability to create any desired nonlinear characteristic, as has been shown experimentally.[2] One possible application of nonlinear demodulation would be the generation of an *equidensity line image*.[3] That is to say, in the *image* $I_B(x,y)$ appear sharp lines, representing a contour map

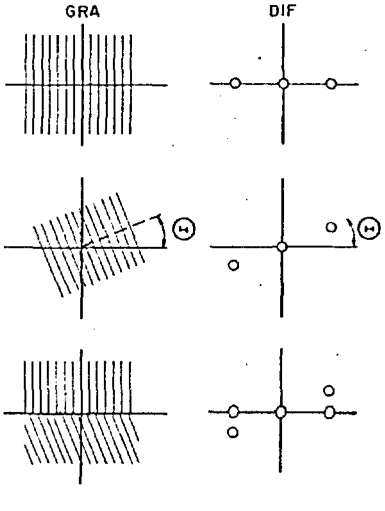


Fig. 3. Three objects with corresponding Fraunhofer diffraction pattern.

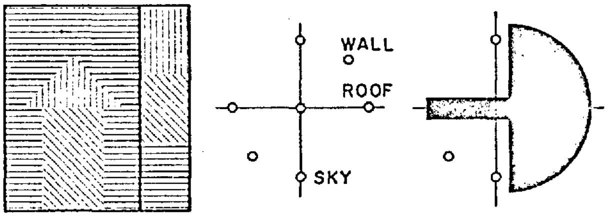


Fig. 4. Modulated object, diffraction pattern, and demodulation mask.

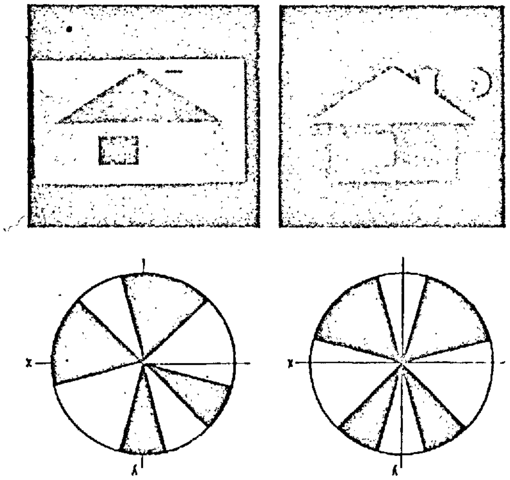


Fig. 5. Result of demodulation procedure for one modulated object with two different demodulation masks.

of the *intensity mountains* of the object $I_0(x,y)$. Representative characteristic curves for modulation and demodulation, as well as the corresponding mask, are shown in Fig. 8. The nonlinear demodulation process has particular significance in the field of optical data processing. Such operations as logic connections, noise suppression, and amplitude quantization can be performed by this technique.

IV. Multiplex Storage

Multiplex transmission in electronics means that more than one message is transmitted at the same time over one channel. By multiplex storage in optics, we mean that more than one image is stored in one place. Multiplex transmission of several messages requires that they be modulated or encoded properly. There are three types of multiplex transmission which can be *translated* into optical multiplex storage.

The best known multiplex method employs different carrier frequencies for each of the individual messages. If one translates the different carrier frequencies into different angular grating orientations, then one arrives essentially at the second color demodulation method of the previous section (Fig. 6). Obviously, one can consider the three color components of the color image as three independent *messages*. Then one might say that the color signal in its theta-modulated form is an example of threefold multiplex storage.

The second multiplex transmission method employs an encoding method where each continuous message is *sampled* at equidistant points and reduced to an equidistant sequence of spikes. The amplitudes of these spikes are representative of the amplitude of the original message within the corresponding sampling interval. All messages are sampled by the same interval, then interlaced and so multiplex-transmitted. The finer the spike width compared with the sampling interval, the more messages can be multiplex-transmitted. This multiplex transmission has been translated already into optical multiplex storage in high-speed photography.[4] Another way to translate interlace multiplexing would be to theta-modulate all image elements which belong to the same signal by the same angle theta. This we have described essentially in our first color demodulation scheme (Fig. 6). All *green* subelements in Fig. 6 constitute together the green signal, which, in principle, can be completely independent from signals which are represented by subelements with different theta angles.

The third type of multiplexing is applicable for digital signals. Here we will consider only binary signals, because of their importance in the field of data processing. First, we will treat this type of multiplexing in mathematical terms, and then a simple experiment will be described.

Let us assume N binary signals

$$S_n(x,y) = [1 \text{ or } 0]; \quad n = 0,1,2,\ldots,N-1$$

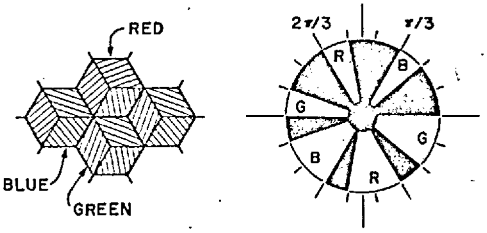


Fig. 6. Color encoding by theta modulation. Each hexagonal object element consists of three gratings, assigned to the three color components here, four object elements: blue, green, red, and blue-green. The demodulation mask consists of transparent and opaque sectors, superimposed by color filters.

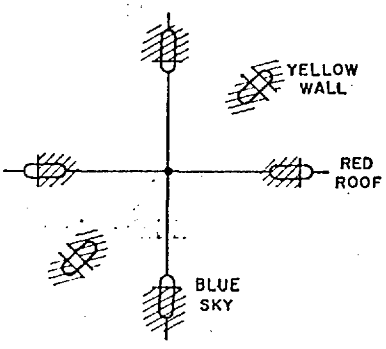


Fig. 7. Color demodulation. The radial structure of the demodulation mask is responsible for color generation, assuming polychromatic light.

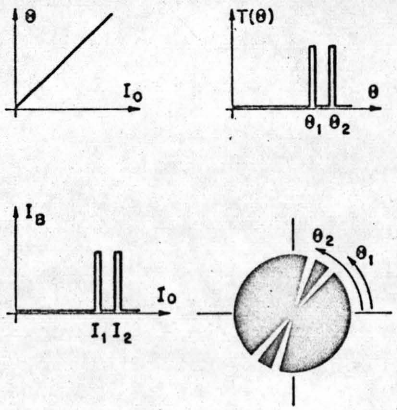


Fig. 8. Equidensity process. Linear modulation $I_0 \rightarrow \theta$; demodulation mask $T(\theta)$ with two small sectors, generating equidensity lines in the *image* I_B at levels I_1 and I_2.

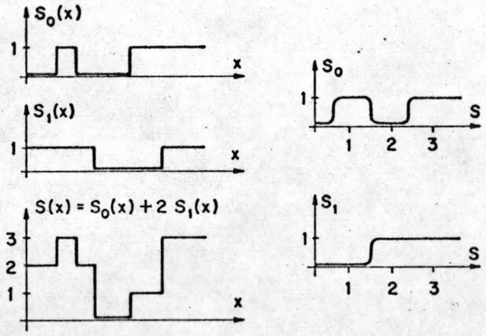


Fig. 9. Multiplex principle. Two signals $S_0(x)$ and $S_1(x)$ are combined to $S(x) = S_0(x) + 2S_1(x)$. Extraction of either S_0 or S_1 is possible by nonlinear process of characteristic $S_0(S)$ or $S_1(S)$.

which we want to store in multiplex fashion. We add them, but with special assigned weighting coefficients 2^n, so

$$S(x,y) = \sum_{0}^{N-1} 2^n S_n(x,y) = 0 \text{ or } 1 \text{ or } 2 \text{ or}, \ldots, \text{or } 2^n - 1.$$

If one wants to extract from the multiplex signal, $S(x,y)$, a particular signal, $S_m(x,y)$, one has to apply a nonlinear characteristic curve of the form $S_m(S)$:

$$S_m(S) = \begin{cases} 1 \text{ if } 2^m(1 + 2p) \leq S < 2^{m+1}(1 + p) \\ 0 \text{ otherwise; } p = 0,1, \ldots, 2^{N-m-1} - 1. \end{cases}$$

This principle is explained for the special case of only two signals ($N = 2$) by means of Fig. 9. The two signals, $S_0(x)$ and $S_1(x)$, the multiplex signal, $S(x) = S_0(x) + 2S_1(x)$, and finally the two nonlinear characteristic curves, $S_0(S)$ and $S_1(S)$, are shown. These nonlinear characteristic curves allow us to extract either S_0 or S_1 from the multiplex signal S, as can be seen easily by inspection of Fig. 9.

To implement these nonlinear characteristic curves, we can use theta demodulation. As original signals, let us assume the symbols AIP and OSA:

$$S_0(x,y) = \begin{cases} 1 \text{ if AIP white} \\ 0 \text{ if AIP black} \end{cases} \Bigg\| \; S_1(x,y) = \begin{cases} 1 \text{ if OSA white} \\ 0 \text{ if OSA black.} \end{cases}$$

The multiplex signal $S(x,y) = S_0(x,y) + 2S_1(x,y)$ obviously contains the amplitude levels 0, 1, 2, and 3. Let us theta-modulate this multiplex signal into $I_M(x,y)$ by assigning to each point (x,y) an angle theta of the grating carrier

$$\theta(x,y) = \tfrac{1}{4}\pi S(x,y).$$

In Fig. 10(a), one sees a photograph of the theta-modulated multiplex signal $I_M(x,y)$ together with its Fraunhofer pattern. If one wants to extract either of the two signals, one has to put $I_M(x,y)$ into the optical demodulation device at M in Fig. 2 and to introduce appropriate demodulation masks. Such masks together with the extracted individual signals are seen in Figs. 10(b) and 10(c). It is common to judge the quality of any multiplex device by the degree of absence of *crosstalk* between different messages after extraction. As Figs. 10(b) and 10(c) show, there is little crosstalk between AIP and OSA.

V. Potential Advantages

Does the concept of theta modulation lead to increased recording density in terms of bits per area? Since the maximum number of bits per field depends only on the system transfer function, on the noise, and on the H and D curve,[5] only practical, rather than fundamental, advantages can be expected. To put this in the proper prospective, let us examine the usefulness of multiplex telephony. If a certain channel

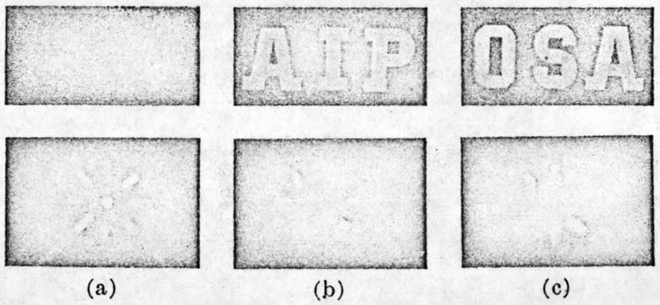


Fig. 10. Multiplex storage. (a) Two binary signals multiplexed and theta-modulated; their Fraunhofer diffraction spectrum; (b), (c) extraction of the two signals, together with corresponding demodulation masks.

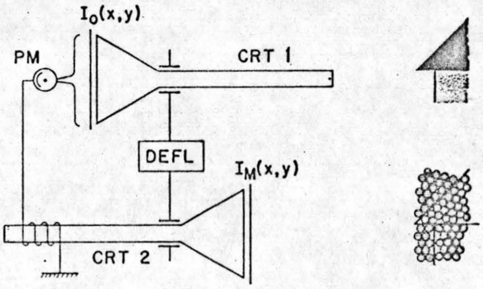


Fig. 11. Modulation process. CRT scanning of object $I_0(x,y)$. Second CRT with grid source. Rotation (theta) of grid images as parts of I_M by means of yoke. Yoke current proportional to photoelectric signal from first CRT. Synchronous and pulsed deflection of both CRT's and synchronously blanked grid source.

has a bandwidth of 5000 cps, but each signal message extends only over 500 cps, then ten signals can be transmitted simultaneously side by side in the frequency domain. The properly combined ten signals are matched to the capacity of the channel. This is one example of optimal adaption of the signals to the channel, but it is not the only way. For example, one could have recorded the signals on tape and then played it back at ten times higher speed. Now, each signal alone fills the whole frequency channel, but for only one tenth of the time. Hence, the total time required is the same, whether one transmits the ten signals sequentially in compressed form, or simultaneously in multiplex-modulated form. In principle, multiplexing does not produce a gain in terms of bits per seconds. Nevertheless, multiplexing is often preferable for practical reasons, as in the case where two-way conversation is wanted.

With the realization that modulation in general can provide only practical, rather than fundamental, advantages, we can apply the above example to an evaluation of theta modulation multiplex storage. Suppose we have one piece of film and four signals, each signal represented by an area equal to that of the given piece of film. Is it more advantageous to reduce these four signals in size such that they will fit side by side on the given piece of film? Or, can we do better by applying theta modulation multiplex storage without reduction in size? In principle, the two approaches are equally good, except for possible practical advantages if only one-to-one reproduction is required. Here, the avoidance of reduction and subsequent reenlargement can be a practical advantage in terms of the required resolution. It could also be an advantage in terms of the energy per unit area in the object plane which is required to produce an enlarged image of adequate illuminance. In general, the photooptical components required to produce the demodulated image need not be capable of resolving the spatial *carrier* frequency, but need only have an entrance pupil large enough to accommodate the diffraction orders which contribute to the final image.

Theta modulation is a new concept in optical data processing which might help to match a given set of signals in a convenient way to the capacity of a given lens or to a given piece of film.

Appendix: The Procedure of Modulation

In terms of the optics, as well as from the viewpoint of processing information, the most significant aspects of theta modulation take place in connection with the demodulation process. We, therefore, simulated the modulation process by photographically recording pieces of finely ruled paper, properly oriented. However, we want to point out at least one modulation technique which leads to a practical implementation.

Consider two synchronously deflected cathode ray tubes (CRT), one for scanning the original object $I_0(x,y)$, the other for printing the theta-modulated signal $I_M(x,y)$. The latter CRT has a multiaperture grid source rather than a point source. Hence, grids of circular extension will be printed on the film I_M. The angular orientation, theta, of this grid pattern is controlled by the current in a yoke around CRT 2 by conventional electron-optical techniques. This current is controlled by the output of the photomultiplier in front of CRT 1 and, hence, is proportional to the transmission of the original object $I_0(x,y)$ at the location of the scanning spot. At the right side of Fig. **11** is shown a portion of an object I_0 *encoded* into its theta-modulated signal I_M. The deflection in both CRTs proceeds stepwise rather than continuously, and the grid source is pulsed synchronously and blanked during deflection.

It is our pleasure to acknowledge many stimulating discussions with R. H. Kay and B. Morgenstern.

References

1. W. E. Glenn, J. Opt. Soc. Am. **48**, 841 (1958); J. Appl. Phys. **30**, 1870 (1959).
2. A. Lohmann and B. Morgenstern, Optik **20**, 450 (1963).
3. E. Lau and W. Krug, *Die Aequidensitometrie* (Akademie-Verlag, Berlin, 1957).
4. J. S. Courtney-Pratt, J. Soc. Motion Picture Television Engrs. **72**, 876 (1963).
5. H. J. Zweig, G. C. Higgins, and D. L. MacAdam, J. Opt. Soc. Am. **48**, 926 (1958).

PROCEEDINGS OF THE CONFERENCE

ON

OPTICAL INSTRUMENTS AND TECHNIQUES

LONDON 1961

EDITOR

K. J. HABELL

M.SC., A.R.C.S., D.I.C., F.INST.P.

National Physical Laboratory, England

LONDON : CHAPMAN AND HALL LTD : 1962

Mock Interferometry

L. MERTZ, N. O. YOUNG AND J. ARMITAGE
Block Associates, Inc., U.S.A.

Summary—A simulated interferometer for Fourier transform spectrometry is described with applications to diverse spectral regions (including ultra-violet) and to emission or absorption line profile studies. In addition to the usual advantages of interference spectrometry, i.e. Fellgett's multiplex advantage(1) of measuring all the colours simultaneously (applicable only in the infra-red) and large acceptance angle at large aperture, i.e. large throughput, the "mock interferometer" has the following three advantages:

1. The dimensional tolerances required are low, making it applicable to the ultra-violet as well as making it extremely durable under adverse operating conditions.
2. It has no beam splitter problem and is applicable to any spectral region for which dispersers are available.
3. The fringe frequency is not necessarily proportional to the radiation frequency. Therefore, by shifting the zero fringe frequency to the neighbourhood of a small spectral region, only a low order (few harmonics) Fourier transformation need be used.

Use of the "mock" interferometer does not, of course, eliminate the need for a Fourier transform computation to obtain the spectrum. This Fourier transformation may be performed with either digital or analogue techniques.

Résumé—Un interferomètre simulé pour spectroscopie par la transformation de Fourier est decrit avec applications à diverses regions spectrales (l'ultraviolet inclus) et aux études de lignes d'emission ou d'absorption spectrales. Aux avantages conventionels de spectroscopie interferentiel; c'est a dire, les avantages multiplex de Fellgett, grace à la mesure simultanée de toutes les couleurs (possible seulement dans l'infrarouge) et la possibilité d'avoir une grande ouverture, l'interferomètre simulé présente de plus les 3 avantages suivants:

1. Les tolérances nécessaires sont faibles. De sorte que l'interferomètre est utilisable dans l'ultraviolet et est très stable même dans les conditions d'opération les plus rudes.
2. Il n'y a pas de problem de separatrice et le principe s'applique à toutes les regions spectrales pour lesquels il y a des disperseurs.
3. Les fréquences des franges ne sont pas nécessairement proportionelles aux fréquences de la lumière. Ainsi l'on peut placer la fréquence zéro des franges au voisinage d'une petite région spectrale et ainsi peut utiliser une transformation de Fourier de bas ordre (contenant peu d'harmoniques).

L'utilization de l'interferomètre simulé n'elimine cependant pas la necessite de calculation de la transformation de Fourier. On peut accomplir cette transformation par les methodes bien connus analogues ou chiffrées.

Zusammenfassung—Es wird ein artaehnliches Interferometer fuer Fourier-transform Spektrometrie mit Anwendungen auf verschiedenen Spektralgebieten (auch im U.V.) beschrieben. Das Interferometer kann zum Studium von Profilen der Emissions oder Absorptionslinien benutzt werden.

Ausser den normalen Vorteilen der Interferenz Spektrometrie (d.h. Felgetts Multiplex Vorteil), dass man alle Farben zur gleichen Zeit misst, (dieser Vorteil bilt nur im

Ultra Rot), und grosse Eintrittswinkel (Durchgang) erreicht, hat das Mock Interferometer die folgenden 3 Vorteile:

1. Die benoetigten Toleranzen sind niedrig, darum kann man das Instrument im U.V. anwenden und umsomehr ist es auch unter widrigen Umstaendensehr dauerhaft.
2. Das Problem der Strahlentrennung besteht nicht, und man kann das Instrument in jedem Spektralgebiet, wo dispergierende Elemente erhalten sind, benutzen.
3. Die Frequenz der Interferenzlinien ist nicht unbedingt proportional zu der Strahlungsfrequenz. Deshalb braucht man nur eine Fouriertransformation niedriger Ordnung, wenn man die Nullfrequenz zerlegt in die Umgebung eines kleinen Spektralgebiets.

Die Notwendigkeit einer Fouriertransformberechnung zur Erhaltung eines Spektrums wird aber nicht von dem Mock eliminiert. Diese Fourier transformation kann entweder mit digital oder mit analog Methoden durchgefuehrt werden.

We call Mock Interferometry the simulation of the channel spectrum transmission, or Edser-Butler bands, of an interferometer. The idea is that if we reproduce such a transmission, regardless of how, we shall indeed have an instrument performing like an interferometer.

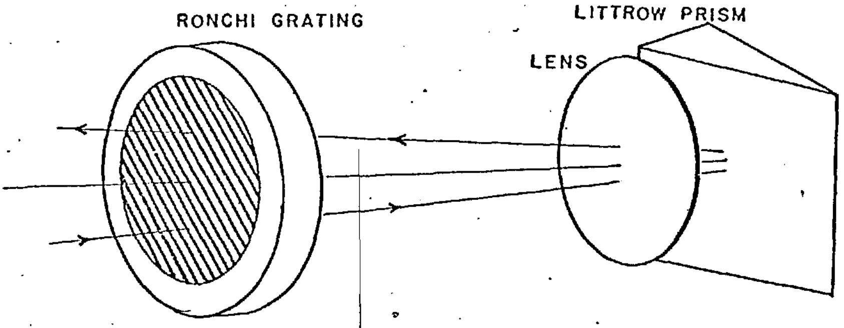


FIG. 1. Mock interferometer.

This transmission is achieved by the straightforward approach of placing a mask over the spectrum formed by a conventional spectroscope. The appropriate mask is clearly a uniformly spaced grill; a Ronchi grating. Now inasmuch as we are taking the overall light transmission through the grill and inasmuch as the grill has uniform spacing, it becomes possible to replace the entrance slit of the spectroscope with a conjugate grill. In this manner we can let a lot more light through while retaining the spectral transmission characteristics.

For example, folding the system we find the Littrow arrangement illustrated in Fig. 1. The entrance and exit grill are combined. One simply uses different regions of this grill for the entrance and exit bundles.

So far, only a single channel spacing has been mentioned. Complete simulation of an interferometer with variable path difference requires variable spacing. Otherwise we would be unable to scan fringes. This is the purpose of the rotating mount in the figure. When the grill is oriented with its lines parallel to the dispersion then the transmission depends on whether the grill is exactly imaged back on itself or with a slight shear. The white or black suggests the zero order transmission of a Michelson interferometer.

This resemblance was experimentally confirmed within ten minutes after its conception. We found that the fringes produced were indistinguishable by eye from those of a Michelson interferometer, and that we could readily scan through white light fringes by rotating the grill. Those of you who have ever sought white light fringes with a Michelson interferometer will appreciate the effortless achievement of white light fringes with even a primitive "mock" interferometer.

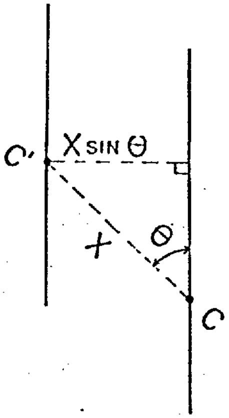


FIG. 2. Abbreviated diagram of entrance-exit parameters.

In order to develop a quantitative knowledge of the transmission, we first notice that the lines of the image of the grill for any colour are necessarily parallel to the lines of the grill itself. In other words, we have one large Moiré fringe. Only the lateral position of the image with respect to the grill determines the transmission or the phase of the Moiré fringe.

In Fig. 2, the pertinent features of the grill system are illustrated. This is a view through the grill, looking down the optic axis. *C* is the centre of rotation of the grill, and *C′* is a monochromatic image of *C*. These points are stationary. There is a line shown through *C* representing the line of the grill which intersects *C*. There is a corresponding parallel image line through *C′*. With x as the distance CC', and θ as the angle between CC' and the grill line through *C*, we find the shear of the image with respect to the grill to be $x \sin \theta$.

The overall Moiré transmission determined by this shear may be expressed

$$T = \frac{1}{2} \cos^2 \left(2\pi \frac{x \sin \theta}{s} + \phi\right),$$

where s is the grill spacing and ϕ is a constant. ϕ is determined solely by the position of *C* on the entrance grill. If the centre of rotation, *C*, lies precisely on the center of a grill line, $\phi = 0$. If *C* lies on the edge of a grill line, then $\phi = 1/2\pi$. It should also be pointed out here that the points *C* and *C′* need not lie within the field of view.

We now make x a function of colour, as a result of the dispersion. In simplest linear form, $x = b(\nu - \nu_0)$, where b is constant and ν_0 is the colour for which C' lies on C. Next we choose a parameter $\tau = \sin\theta$, and the transmission becomes

$$T = \frac{1}{2}\cos^2\left[2\pi\frac{b(\nu - \nu_0)\tau}{s} + \phi\right]$$

When $\phi = C_1$ and $\nu_0 = 0$, this is the transmission of a Michelson interferometer at retardation $b\tau/s$.

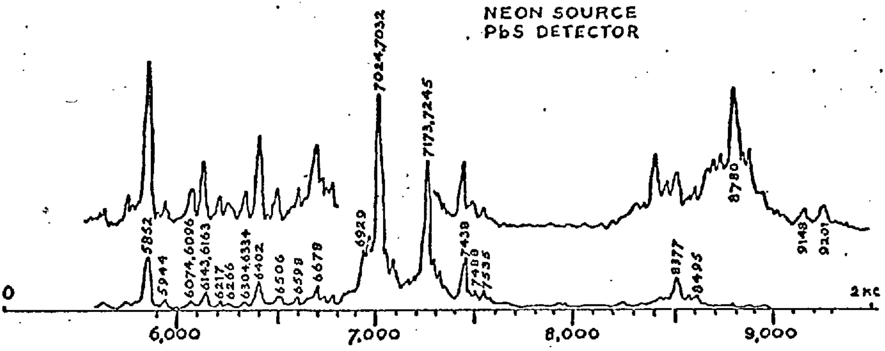


FIG. 3. Neon spectrum obtained with mock interferometer.

Nothing serious happens with non-linear dispersion. The wavenumber need not actually represent radiation frequency, but a colour scale linearized to the dispersion. Measurements in terms of this colour scale may later be calibrated into terms of the original radiation frequency. The same type of wavelength calibration is required of all prism spectrometers.

If the dispersing element of the Littrow spectrometer is a diffraction grating, then we find that the fringe frequency ν is linear with wavelength (for small dispersion angles) as will be illustrated shortly.

The operation of the instrument proceeds as with systems of Fourier transform spectrometry involving Michelson interferometers. These techniques have found increasing use ever since their advantages were first realized by Fellgett and most of the details were presented at the Paris conference on Interference Spectroscopy in 1957, and at the Teddington conference on Interferometry in 1959.

The resolving power using the "mock" interferometer clearly cannot exceed the resolving power of the component Littrow spectrometer. For maximum resolving power, the grill should be as fine as the spectrometer will resolve. If the grill were made finer than this, no fringes would occur, since the original Ronchi grating is not resolved.

A preliminary spectrum obtained with our "mock" interferometer is shown in Fig. 3. This shows a neon spectrum obtained from our interferogram with fringe frequencies up to two kilocycles. Notice that the wavelength scale is linear, that long wavelengths have high fringe frequencies, and that zero fringe frequency lies near 5000 Å. This latter ability to locate zero fringe frequency in the spectrum at will, allows us to achieve higher resolution than the size of our Fourier computation would normally permit. We expect this ability to be one of the fundamental merits of "mock" interferometry.

The instrument with which the neon spectrum was obtained is illustrated in Fig. 4. Six seconds were used for recording signal on magnetic tape, and the source was a small neon lamp of about 1/4 watt.

The principle engineering problem is the construction of the drive such that sin θ is linear with time. So far, we have managed two approaches to the problem. The first approach was a rapid repetitive scan ($2\frac{1}{2}$ scans per second) in order to make the output compatible with magnetic tape recording and audio frequency wave analysis.$^{(2)}$ This involves extremely non-uniform rotation and the accelerations and backlash prevented operation of a cam corrected drive. Instead, an approximate drive was made by using a crank and slot connection between offset shafts. It turns out to be important that the slot drive the crank, rather than vice versa, in order to approximate the desired motion. With this system we are able to use a duty cycle of about $\frac{1}{4}$, blanking the amplifier during the remaining $\frac{3}{4}$ time, and we get about $\frac{1}{2}$ the maximum resolution. By that we mean we used the region $-\frac{1}{2} < \sin\theta < +\frac{1}{2}$ for the measurement.

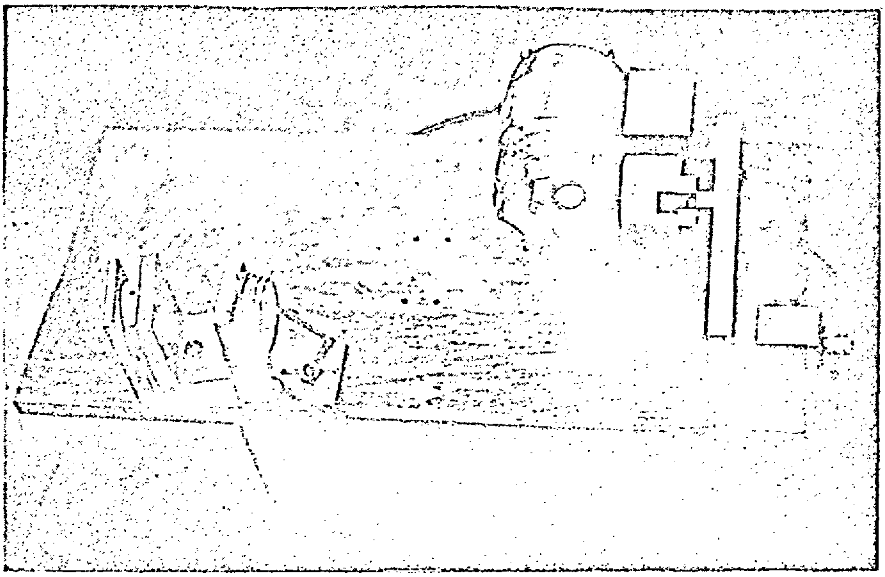


FIG. 4. Mock interferometer.

Another drive which we have recently constructed for the visible and ultraviolet, where we don't have low frequency detector noise and so can scan slowly, is an escapement with non-uniformly spaced teeth. The Fourier transformation in this case is readily adaptable to digital computation.

In conclusion, we would like to mention some of our desired applications of "mock" interferometry. The first is high resolution photoelectric stellar spectrophotometry. Although Fellgett's multiplex gain is on the average balanced by the increased photon noise, the throughput gain is still available. With conventional spectrophotometers it is impossible to decrease the slit width to gain resolution simply because the star image is too big.

We would also like to apply "mock" interferometry to the vacuum ultraviolet, not only for low resolution work but also for high resolution study of the Lyman α profile.

It has recently come to our attention that in 1959, Lohmann[3] mentioned the possibility of application of Moiré fringes to spectral analysis. As has been seen, the "mock" interferometer also employs Moiré fringe concepts although in a different way.

Finally, we would like to express our appreciation to the Geophysics Research Directorate for their support of this research.

REFERENCES

(1) Fellgett P., *J. Physique Rad.*, 19, 187 (1958).
(2) Mertz L., *J.O.S.A.*, Advmt., March, 1960.
(3) Lohmann A., *Optica Acta*, 6, 37 (1959).

DISCUSSION

Perry: I would like to ask Dr. Mertz to what extent he intends to press the resolution aspects. I gathered that he needed a grating of a frequency as high as would be justified by the resolving power. I can foresee quite interesting design problems here; it would seem to me that he would need an anastigmatised, achromatised objective which would still leave outstanding the effects of curvature of spectral lines, which over an appreciable area might introduce some reduction of resolving power. If it is intended to be pressed so far, an achromatised Schmidt system might be desirable, but you cannot correct curvature of the lines.

Mertz: I would like to say this. The curvature of the lines gives fringes which correspond to the circular fringes of a Michelson interferometer, and before we get to the circular fringes we still have an appreciable advantage over a conventional spectrometer. One can avoid the curvature of the spectrum lines by using a combination prism grating that Connes has described—I believe he uses it in SISAM—that is free from curvature of the spectrum lines. We would like to do high resolution studies, we would like to do studies of Lyman α profile, and there are a lot of stellar absorption line profiles that I am anxious to get at.

Perry: I was only thinking in terms of the system that you projected on the screen of course.

Mertz: Well one can use any sort of disperser diffraction grating.

Perry: Shouldn't one understand that any reduction in curvature of spectral lines means a reduction in dispersion?

Mertz: No, one can compensate by using a prism and grating combination, so that as you look at the grating it appears rather flat.

Perry: What about the question of anastigmatism?

Mertz: Astigmatism is a serious problem we have not been able to get round. We would like to use the system with an Ebert spectrometer but astigmatism so far has prevented that application.

Schawlow: Have you given any thought to applying mock interferometry to real interferometry, using something other than a prism as a dispersive element?

Mertz: Yes, in fact we have given thought to using some scatter screens which Jim Burch was kind enough to show us, but that would be too lengthy to talk about right here--that makes it more of a monochromator.

RING: It may be of interest to comment that we have been using the slit system of your interferometer for an analogue Fourier transformer by taking two line-screens (200 lines/inch), putting them at a slight angle and putting these over a mask of an interferogram from a Michelson interferometer, or from your system.

MERTZ: Do you get any good results from this sort of application?

RING: So far we have been working it both ways, by feeding into the light source a delta function, making a mask, and then transforming this and getting a delta function back. Both screens have to be tilted, and the Moiré fringes scanned across the mask, and so we get a sequential Fourier transform. Although we have been working on this for some time (since we saw your abstract), it was a long time before we saw how to use it this way. For high resolution, at least three lines of the line-screen must fill a spacing of the fringe to get a well defined moiré fringe.

MERTZ: Let me be a little more clear, by high resolution, I do not mean resolution of the order of 500,000, I mean resolution of less than 50,000. In astronomy this is considered highly dull resolution.

RING: How do you propose to make the line-screens?

MERTZ: By a photo-engraving technique—they do not have to be self-supporting.

MARÉCHAL: A similar device was described by Girard: what is the difference?

MERTZ: The difference between this device and Girard's is that this uses the Fourier transform technique and in so doing gains Fellgett's advantage as well as the wide field advantage and one has to have Fellgett's advantage to balance photon noise in the visible and the ultra-violet. If you just have a selective modulator you have lost in the visible or the ultra-violet.

MARÉCHAL: Yes, but Girard was against conventional equipment.

MERTZ: Yes, but you also get much more photon noise. You have to gain the advantages of both Fellgett and of acceptance angle.

DITCHBURN: In the vacuum ultra-violet you gain only by increase of light; presumably if one has enough light then one gets as good a spectrum out of an ordinary conventional spectrograph?

MERTZ: Yes. This and a conventional spectrograph are exactly the same on the average for the same amount of light passing through the instrument. However, instead of an entrance slit we can use big entrance apertures, which means that we can pass much, much, much more light through the instrument.

DITCHBURN: Yes but your grid must be very, very fine and it must be made of wires.

MERTZ: But you can also get a very high dispersion in the ultra-violet too, so that the grid does not really have to be all that fine.

DITCHBURN: To get a very high dispersion you must have a very long instrument and that implies a very low f/value and therefore you would have lost your light again!

MERTZ: On the whole we still do better than a lot of conventional instruments, I believe.

INGELSTAM: How does linear dispersion come in? You can use different dispersive elements but you have a scale which is linear in wave number.

MERTZ: The scale does not really have to be of wave number; it can be just some sort of colour scale.

INGELSTAM: That is true, but on the other hand you have a certain intensity in a certain scale, it is given by the dispersion of the prism or the grating, but when you make this transform don't you have to know the final intensity in order to have the Fourier transform linear?

MERTZ: No, you don't have to modify intensity. There is vignetting in the colour, so that you do get cut off at the ends of the spectrum, but that is only due to vignetting and has nothing to do with the dispersion.

INGELSTAM: Even for very long regions of the spectrum?

MERTZ: Yes, if you can get the light to the detector.